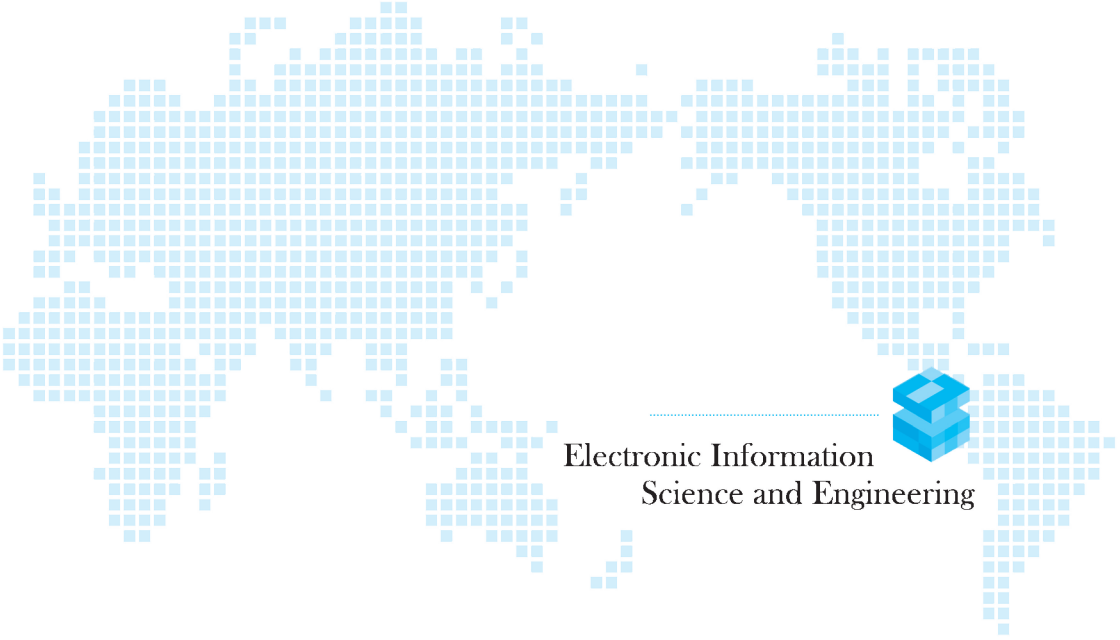
全国电子信息类和财经类优秀教材
普通高等教育“十三五”规划教材

信息论基础与应用

李 梅 编著



電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内 容 简 介

信息论是现代信息通信领域的基础理论，是研究信息传输和信息处理的一般规律的科学。我们在借鉴了国内外众多的信息论优秀教材和参考资料之后编写了本书。本书以香农的三个编码定理为中心，重点讲述了相关的基本概念、原理、方法和应用。本书介绍经典信息论的内容，不涉及过多的分支。全书选择了很多与日常生活密切相关，具有一定趣味性的习题，同时增加了动手编程实践。

本书可作为通信及电子信息类相关专业高年级本科生和研究生的教材，也可作为相关专业科研人员的参考书。

图书在版编目（CIP）数据

信息论基础与应用/李梅编著. —北京：电子工业出版社，2016.6
ISBN 978-7-121-29026-8

Ⅰ. ①信… Ⅱ. ①李… Ⅲ. ①信息论—高等学校—教材 Ⅳ. ①G201

中国版本图书馆 CIP 数据核字（2016）第 128738 号

策划编辑：章海涛
责任编辑：章海涛
印　　刷：
装　　订：
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编 100036
开　　本：787×1 092　1/16　　印张：14.5　　字数：370 千字
版　　次：2016 年 6 月第 1 版
印　　次：2016 年 6 月第 1 次印刷
定　　价：36.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：192910558（QQ 群）。

前　言

信息论是现代信息通信领域的基础理论，是研究信息传输和信息处理的一般规律的科学，因此目前各高等院校的相关专业的本科生、研究生都开设了这门课。

在借鉴了国内外众多的信息论优秀教材和参考资料之后，作者根据多年的教学实践经验编写了本书。本书以使读者掌握基本概念和方法为目的，力图以读者最易接受的方式介绍信息论的基本内容及应用。本书可作为通信及电子信息类相关专业高年级本科生和研究生的教材，也可作为相关专业科研人员的参考书。

教材以香农的三个编码定理为中心，重点讲述了相关的基本概念、基本原理和基本方法。鉴于目前各大高校都在削减学时，教材只介绍经典信息论的内容，没有涉及过多的分支。同时，我们认为，“信息论”是一门理论和实践紧密结合的课程，因此选择了很多与日常生活密切相关，具有一定趣味性的习题，同时增加了动手编程实践。

除绪论外，本书还包括 7 章内容。绪论主要介绍香农信息论的研究对象、目的和内容。第 1 章介绍信息度量的几个重要概念：自信息、互信息、信息熵、平均互信息以及数据处理定理。第 2 章研究定量度量信源产生信息的能力和信源冗余度的问题。第 3 章研究定量描述信道传递信息能力的问题，并介绍了信道容量的计算方法。第 4 章的核心内容是香农的无失真信源编码定理。围绕这个定理我们介绍了无失真信源编码的基本概念，讲述了几种实用的无失真信源编码方法。第 5 章讲述香农的有噪信道编码定理以及纠错编码的主要内容，介绍了信道编码的基本概念、基本理论。第 6 章介绍香农的限失真信源编码定理，引入了信息率失真函数的概念并介绍了信息率失真函数的性质以及计算方法，还介绍了几种常用的熵压缩编码算法。第 7 章介绍信息论的主要应用。

在此感谢李亦农、吴韶波、王永峰、李红、杨师、徐益民、杨世忠和徐安庭等老师提出的宝贵意见，同时感谢邢开颜、赵永平、刘旭、甄晓丹、明舒晴同学为收集习题所做的大量工作。

在本书的编写过程中，参阅了国内外的很多信息论经典著作（列于本书参考文献中），同时参考了大量国内外知名大学信息论课程的课后习题及解答，在此向有关作者表示感谢！

尽管我们在本书中力求更加符合读者的要求，但仍无法避免错漏和不当，恳切希望广大读者提出宝贵意见。欢迎交流探讨：maggieli@ cugb. edu. cn。

李　梅

2016 年 5 月于中国地质大学（北京）

目　　录

绪　　论

本章首先介绍信息理论中最重要的概念——信息，澄清与之容易混淆的几个概念，然后介绍“信息论”课程的学习目的和内容，希望读者对信息论的研究对象、研究内容以及学习信息论的意义有整体的了解。

0.1　信息的概念

信息论是通信的数学基础，它是随着通信技术的发展而形成和发展起来的一门新兴的横断学科。信息论创立的标志是 1948 年 Claude Shannon（香农）发表的论文 *A Mathematical Theory of Communication*。为了解决在噪声信道中有效传输信息的问题，香农在这篇文章中创造性地采用概率论的方法来研究通信中的问题，并对信息给予了科学的定量描述，第一次提出了“信息熵”的概念。

在日常生活中，人们往往对“消息”和“信息”不加区分，消息被认为就是信息。例如，收到一封电报或者听了天气预报，人们就说得到了信息。

收到消息后，如果消息告诉了人们很多原来不知道的新内容，人们会感到获得了很多的信息，而如果消息是人们基本已经知道的内容，那么所获到的信息并不多。所以，信息应该是可以度量的。那么，怎样度量信息呢？人们需要一个可以用数学模型来表示的信息概念。

1928 年，哈特莱（Hartley）首先提出了用对数度量信息的概念。一个消息包含的信息量用其所有可能取值的个数的对数来表示。比如，抛掷一枚硬币可能有两种结果：正面和反面，所以得知抛掷结果后获得的信息量是$\log_2 2 = 1$ 比特。而一个十进制数字可以表示 0 ~ 9 中的任意一个符号，所以一个十进制数字包含$\log_2 10 = 3.3219$ 比特的信息量。这里的对数取以 2 为底，信息量的单位为比特（bit）。

哈特莱的工作给了香农很大的启示，香农进一步注意到，消息的信息量不仅与它的可能值的个数有关，还与消息本身的不确定性有关。例如，抛掷一枚偏畸硬币，如果正面向上的可能性是 90%，当人们得知抛掷结果是反面时得到的信息量，会比得知抛掷结果是正面时得到的信息量大。

一个消息之所以会包含信息，正是因为它具有不确定性，一个不具有不确定性的消息是不会包含任何信息的。通信的目的就是为了消除或部分消除这种不确定性。比如，在得知硬币的抛掷结果前，人们对于结果是出现正面还是出现反面是不确定的，通过通信，人们得知了硬币的抛掷结果，消除了不确定性，从而获得了信息。因此，信息是对事物运动状态或存在方式的不确定性的描述。这就是香农信息的定义。

用数学的语言来讲，不确定就是随机性，具有不确定性的事件就是随机事件。因此，可运用研究随机事件的数学工具——概率来测度不确定性的大小。在信息论中，人们把消息用

随机事件表示，发出这些消息的信源则用随机变量来表示。比如，抛掷一枚硬币的试验可以用一个随机变量来表示，而抛掷结果可以是正面或反面，这个具体的消息则用随机事件表示。

某个消息 x_i 出现的不确定性的大小被定义为自信息，用这个消息出现的概率的对数的负值来表示，即

$$I(x_i) = -\log p(x_i) \tag{0.1}$$

自信息同时表示这个消息包含的信息量，也就是能够给予收信者的最大信息量。如果消息能够正确传输，收信者就能够获得这样多的信息量。

信源包含的信息量定义为信源发出的所有可能消息的平均不确定性。香农把信源包含的信息量称为信息熵。自信息的统计平均定义为信源熵，即

$$H(X) = -\sum_{i=1}^{q} p(x_i)\log p(x_i) \tag{0.2}$$

式中，q 表示信源消息的个数。信息熵表示信源的平均不确定性的大小，同时表示信源输出的消息所含的平均信息量。因此，虽然信源产生的消息可能包含不同的信息量，如抛掷一枚偏畸硬币的结果“是正面”和“是反面”这两个消息所含的信息量不同，但是可以用它们的平均值来表示这个信源（抛掷一枚偏畸硬币的试验）的平均不确定性。

在接收端，信源的不确定性得到了部分或全部消除，接收者就得到了信息。信息在数量上等于通信前后“不确定性”的消除量（减少量）。这种建立在概率模型上的信息概念排除了日常生活中“信息”一词主观上的含义和作用，只是对消息的统计特性的定量描述，所以信息可以度量，而且与日常生活中“信息”的概念并不矛盾，因此是一个科学的定义。根据这样的定义，同样一个消息对于任何接收者来说，包含的信息量都是一样的。事实上，信息具有很强的主观性和实用性，同样一个消息对不同的人常常有不同的主观价值或主观意义。例如，同一则气象预报对在室外工作的人和在室内工作的人，可能会有不同的意义和价值，因此所提供的信息量也应该不同。所以，香农信息在某些情况下也具有一定的局限性。

0.2 信息论的研究对象、目的和内容

从诞生到现在，信息论虽然只有短短的几十年，但它的发展对学术界及人类社会的影响是相当广泛和深刻的。如今，信息论的研究内容不仅包括通信，还包括所有与信息有关的自然和社会领域，如模式识别、计算机翻译、心理学、遗传学、神经生理学、语言学、语义学甚至社会学中有关信息的问题。香农信息论迅速发展成为涉及范围极广的广义信息论——信息科学。

信息论的研究对象是广义的通信系统，它把所有的信息流通系统都抽象成如图 0.1 所示的模型。这个模型不仅包括电话、电报、传真、电视、雷达等狭义的通信系统，还包括生物有机体的遗传系统、神经系统、视觉系统甚至人类社会的管理系统。信息是以消息的形式在这个通信系统中传递的。通过研究通信系统中消息的传输和处理，人们得到信息传输与处理的规律，以提高通信的可靠性和有效性。

任何信息流通系统中都有一个发出信息的发送端（信源）、一个接收信息的接收端（信宿）以及信息流通的通道（信道）。在信息传递的过程中不可避免地会有噪声，所以会有一

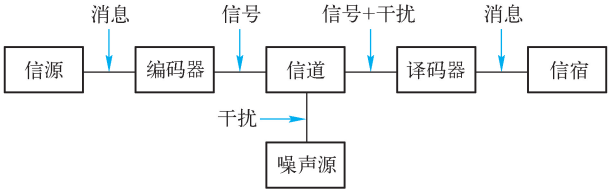


图 0.1　通信系统模型

个噪声源。为了把信源发出的消息变成适合在信道中传输的信号，还需要加入编码器，在送到信宿之前要进行反变换，所以要加入译码器。

这个通信系统主要包括如下 5 部分。

（1）信源

顾名思义，信源是产生消息和消息序列的源。信源可以是人、生物、机器或其他事物。比如，“各种气象状态”是信源，能够产生独特气味吸引蜜蜂采蜜的花朵是信源，人的大脑活动也是一种信源。信源的输出是消息（或消息序列）。

消息有不同的形式，如文字、符号、语言、图片、图像、气味等。消息以能被通信双方所理解的形式，通过通信进行传递和交换。消息携带着信息，是信息的载体。

信源输出的消息是随机的、不确定的，但有一定的规律性，因此用随机变量或随机矢量等数学模型来表示信源。

（2）编码器

编码就是把消息变成适合在信道中传输的物理量，这种物理量称为信号（如电信号、光信号、声信号、生物信号等）。信号携带着消息，它是消息的载体。

编码器可分为信源编码器和信道编码器。

信源编码的目的是压缩信源的冗余度（即多余度），提高信息传输的效率，这是为了提高通信系统的有效性。信源编码又可分为无失真信源编码和限失真信源编码。

信道编码是为了提高信息传输的可靠性而有目的地对信源编码器输出的代码组添加一些监督码元，使之具有纠错、检错能力。比如，老师讲课需要把知识进行加工和提炼，以提高信息传输的有效性，而为了让学生听得明白，有时需要适当重复，这是为了提高信息传输的可靠性。

在实际的通信系统中，可靠性和有效性常常是矛盾的，提高有效性必须去掉信源符号的冗余部分，但是这会导致可靠性的下降；而提高可靠性需要增加监督码元，这又降低了有效性。有时为了兼顾有效性，不一定要求绝对准确地在接收端再现原来的消息，而是可以允许一定的误差或失真，也就是说，允许近似地再现原来的消息。

（3）信道

信道是指通信系统把载荷消息的信号从发送端送到接收端的媒介或通道，是包括收发设备在内的物理设施。除了传播信号外，信道还有存储信号的作用。在狭义的通信系统中，实际信道有明线、电缆、波导、光纤、无线电波传播空间、磁盘、光盘等，这些都属于传输电磁波能量的信道。对于广义的通信系统来说，信道还可以是其他传输媒介。

在信道中引入噪声和干扰是一种简化的表达方式。为了分析方便，系统其他部分产生的

干扰和噪声都被等效地折合成信道干扰，被视为由一个噪声源产生，它将作用于所传输的信号上。这样，信道输出的已是叠加了干扰的信号。噪声源的统计特性是划分信道的依据，并且是信道传输能力的决定因素。因为干扰或噪声往往具有随机性，所以信道用输入和输出之间的条件概率来描述。

（4）译码器

译码就是把信道输出的已迭加了干扰的编码信号进行反变换，变成信宿能够理解的消息。译码器也可分为信源译码器和信道译码器。译码器需要尽可能地再现信源输出的消息。

（5）信宿

信宿是消息传送的对象，即接收消息的人或机器。

以上考虑的是收发两端单向通信的情况，只有一个信源和一个信宿，信息传输也是单向的。在组网通信的情况下，如电话网、计算机网络等，可能有很多单独的信源、信道和信宿同时进行信息交换。例如，广播信道是一个输入、多个输出的单向信道，卫星通信则是多个输入、多个输出的多向传输的通信。这就需要把两端单向通信的模型进行适当修正，得出多用户通信系统的模型，即把两端单向通信的信息理论发展成为多用户通信信息理论。

信息论研究的是关于这个通信系统的最根本、最本质的问题。例如：

① 什么是信息？如何度量信息？

② 怎样确定信源的输出中含有多少信息量？

③ 对于一个信道，它传输信息量的最高极限（信道容量）是多少？

④ 为了能够无失真地传输信源信息，对信源编码时所需的最少的码符号数是多少？（无失真信源编码即香农第一定理）

⑤ 在有噪信道中有没有可能以接近信道容量的信息传输率传输信息而错误概率几乎为零？（有噪信道编码即香农第二定理）

⑥ 如果对信源编码时允许一定量的失真，所需的最少的码符号数又是多少？（限失真信源编码即香农第三定理）

毫无疑问，如果对这些问题都有了确定的答案，那么在设计通信系统时就有了目标和指导方向，也有了评价通信系统优劣的标准。

下面举几个成功地应用信息论的概念和方法指导通信系统设计的例子。

（1）无失真信源编码的应用之一：计算机文件的压缩

由于数据库的广泛应用，存储计算机文件所需的存储量问题日益突出。目前，对计算机文件的压缩已发展了至少 20 多种算法，其中较好的算法能使文件压缩后所需的存储量只为原文件的 30% 左右。

（2）有噪信道编码的应用之一：模拟话路中数据传输速率的提高

最早的调制解调器的速率只有 300 bps，此后调制解调器的速率为 4800 bps、9600 bps、14. 4 kbps、19. 2 kbps、28. 8 kbps，如今的实际速率已达到 33. 6 kbps，非常接近于理论极限。

（3）限失真信源编码的应用之一：语音信号压缩

按照信息理论的分析，话音信号所需的编码速率可以远远低于按奈奎斯特采样定理和量化噪声理论所确定的编码速率。几十年来，人们在这方面的工作取得了巨大进展。CCITT 关于长途电话网的语音编码速率标准已从 1972 年 G. 711 标准中的 64 kbps 降低到 1992 年标准

中的 16 kbps。在移动通信中，1988 年，欧洲 GSM 标准中的语音编码速率为 13.2 kbps。1989 年，美国 CTIA 标准中的速率为 7.95 kbps。目前，声码器的速率可低于 100 bps，已接近信息论指出的极限。

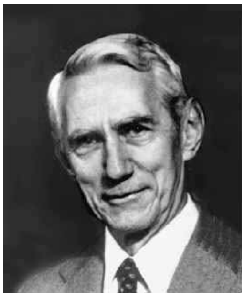
信息论的成就给许多学科带来了希望之光，人们试图应用信息的基本理论解决诸如组织化、语义化、听觉、神经学、生理学、心理学中一些难以解决的问题。目前，人们已将信息论广泛应用于物理、化学、生物学、心理学、管理学等学科，信息的概念和方法已广泛渗透到各个科学领域，迫切要求突破香农信息论的狭隘范围，以便推动许多新兴学科的进一步发展。一门研究信息的科学——广义信息论正在形成。对于信息论的研究范围，一般有如下 3 种理解：

① 狭义信息论：又称为香农信息论，主要通过数学描述和定量分析，研究通信系统从信源到信宿的全过程，包括信息的测度、信道容量以及信源和信道编码理论等问题，强调通过编码和译码使收、发两端联合最优化，并且以定理的形式证明极限的存在。这部分内容是信息论的基础理论。

② 一般信息论：也称为工程信息论，主要研究信息传输和处理问题，除香农信息论的内容外，还包括噪声理论、信号滤波和预测、统计检测和估计、调制理论、信息处理理论、保密理论等。

③ 广义信息论：不仅包括上述两方面的内容，还包括所有与信息有关的自然和社会科学领域，如模式识别、机器翻译、心理学、遗传学、神经生理学、语言学、语义学甚至社会学中有关信息的问题。

扩展阅读： 信息论的形成和发展



香农

克劳德·艾尔伍德·香农（Claude Elwood Shannon，1916—2001 年），美国数学家、信息论创始人。1948 年和 1949 年，香农在《贝尔系统技术杂志》（*Bell System Technical Journal*）上分别发表了著名论文《通信的数学原理》和《噪声下的通信》。香农阐明了通信的基本问题，给出了通信系统的模型，提出了信息量的数学表达式，并解决了信道容量、信源统计特性、信源编码、信道编码等一系列基本问题。这两篇论文成为信息论的奠基性著作。

从有人类的那一天开始，人类就生活在信息的海洋里。人类日常生活、工农业生产、科学研究和战争等，一切都离不开消息传递和信息流动。为了突破语言、文字的局限性，人类不断地创造出许许多多的信息传递方法。1844 年，美国人莫尔斯发明了高效率编码电报法；1876 年，贝尔发明了世界上第一台可用的电话机；1887 年，马可尼发明了无线电报，加上调幅广播、电视、调频广播、数字通信系统、声码器、扩频通信等，通信系统已日益成为人类社会的神经系统。

信息论是在长期的通信工程实践和理论研究的基础上发展起来的，它研究如何认识信息，利用信息，以改变自己的生存条件、创造更好的生活环境等。尽管人类对信息的认识、利用源远流长，但真正对信息理论的研究只有半个多世纪。

1922 年，卡松提出了边带理论，指明了信号在调制（编码）和传送的过程中与频谱宽

度的关系。1924 年，美国的奈奎斯特证实了卡松的理论。1928 年，哈特莱发表了《信息的传输》，首先提出了消息是代码、符号、序列，而不是内容本身。他第一次提出了“信息量”的概念，并试图用数字公式加以描述，为信息论的创立提供了思路。1945 年，莱斯对噪声的研究做了全面的总结，通信理论已经全面走上统计分析之路。1946 年计算机和 1947 年晶体管的诞生与相应技术的发展，则是信息论产生的物质基础。

第二次世界大战中，由于通信在军事上的重要意义，香农开始从事信息论的研究。1948 年，香农发表了《通信的数学理论》，主要内容是研究信源、信宿、信道及编码问题。战后，由于通信事业的需要和电子技术的飞速发展，促进了信息论的进一步发展，许多国家的学者对此进行了大量的研究工作，并卓有成就。1951 年，美国无线电工程学会承认了信息论这门新学科，建立了信息论学组。

扩展阅读： 量子通信与量子信息论

20 世纪 80 年代，量子力学与信息科学相结合，诞生了一门新型的交叉学科——量子信息学（Quantum Information Theory），主要包括量子通信和量子计算，为确保信息安全和提高计算速度提供了全新的方案。

量子通信主要基于量子纠缠态的理论，使用量子隐形传态（传输）的方式实现信息传递。根据实验验证，具有纠缠态的两个粒子无论相距多远，只要一个发生变化，另外一个也会瞬间发生变化。利用这个特性实现量子通信的过程如下：事先构建一对具有纠缠态的粒子，将两个粒子分别放在通信双方，将具有未知量子态的粒子与发送方的粒子进行联合测量（一种操作），则接收方的粒子瞬间发生坍塌（变化），坍塌（变化）为某种状态，这个状态与发送方的粒子坍塌（变化）后的状态是对称的，然后将联合测量的信息通过经典信道传送给接收方，接收方根据接收到的信息，对坍塌的粒子进行幺正变换（相当于逆转变换），可得到与发送方完全相同的未知量子态。

量子纠缠可以用“薛定谔猫”来帮助理解：把一只猫放到一个有毒的盒子中，然后将盒子盖上，然后问这只猫现在是死了还是活着。量子物理学的答案是：它既是死的，也是活的。有人会说，打开盒子看一下不就知道了。是的，打开盒子确实能知道猫是死是活，但按量子物理的解释：这种死或活的状态是人为观察的结果，也是人的宏观干扰使得猫变成了死的或活的，并不是盒子盖着时的真实状态。同样，微观粒子在不被“干扰”之前就一直处于“死”和“活”两种状态的叠加，也可以说它既是“0”也是“1”。

以笛卡儿、伽利略、牛顿为代表的主流物理学家认为，宇宙的组成部分相互独立，它们之间的相互作用受到时空的限制。而量子纠缠效应脱离了时空，证实了任何两种物质之间，不管距离多远，都有可能相互影响，不受四维时空的约束，是非局域的（nonlocal），不仅宇宙证实了“爱因斯坦的幽灵”——超距作用（spooky action in a distance）的存在，也证实了中国人一直强调的因果报应等观点——任何两种物质在冥冥之中存在深层次的内在联系。

与量子通信相比，经典通信的安全性和高效性都无法与之相提并论。量子通信绝不会“泄密”，确保了通信的安全性。其一，量子加密的密钥是随机的，即使被窃取者截获，也无法得到正确的密钥，因此无法破解信息；其二，通信双方分别有两个处于纠缠态的粒子，其中一个粒子的量子态发生变化，另一方的粒子量子态就会立刻随之变化。根据量子理论，

宏观的任何观察和干扰都会立刻改变量子态，导致其坍塌，因此窃取者由于干扰而得到的信息已经破坏，并非原有信息。高效性体现在被传输的未知量子态在被测量之前会处于纠缠态，即可以同时代表多个状态。例如，一个量子态可以同时表示0和1两个数字，7个这样的量子态就可以同时表示128个状态或128个数字：0～127。量子通信的这样一次传输就相当于经典通信方式的128次。可以想象，如果传输带宽是64位或更高，那么效率之差将是惊人的。

量子通信是国际量子物理和信息科学的研究热点。我国从20世纪80年代开始从事量子光学领域的研究，并且在量子通信领域开始领跑世界。

1997年，在奥地利留学的中国青年学者潘建伟与荷兰学者波密斯特等人合作，首次实现了未知量子态的远程传输。这是国际上首次实验成功将一个量子态从甲地的光子传输到乙地的光子上。实验中传输的只是表达量子信息的“状态”，作为信息载体的光子本身并不被传输。

2006年夏，中国科学技术大学教授潘建伟小组、美国洛斯·阿拉莫斯国家实验室、欧洲慕尼黑大学－维也纳大学联合研究小组，各自独立地实现了诱骗态方案，同时实现了超过100 km的诱骗态量子密钥分发实验，由此打开了量子通信走向应用的大门。

2008年底，潘建伟的科研团队成功研制了基于诱骗态的光纤量子通信原型系统，在合肥成功组建了世界上首个3节点链状光量子电话网，成为实用化量子通信网络实验研究的两个团队之一（另一个团队为欧洲联合实验团队）。

2009年9月，潘建伟的科研团队在3节点链状光量子电话网的基础上，建成了世界上首个全通型量子通信网络，首次实现了实时语音量子保密通信，标志着中国在城域量子网络关键技术方面已经达到了产业化要求。

2012年，中国科学家潘建伟等人在国际上首次成功实现百千米量级的自由空间量子隐形传态和纠缠分发。

2016年8月16日，由潘建伟担任首席科学家的世界第一颗量子科学实验卫星“墨子号”顺利发射升空。“墨子号”是中国科学院空间科学先导专项首批实验卫星之一，主要科学目标是星地高速量子密钥分发实验，在此基础上实验广域量子密钥网络，以期空间量子通信实用化；在太空中分发纠缠光子，实验量子隐形传态，并检验空间尺度的量子力学完备性。2016年年底，“京沪干线”将建成，全长超过2000 km，将成为连接北京、济南、合肥、上海等城域网络的量子保密通信线路，也将是全球首个远距离广域光纤量子保密通信骨干线路。

第 1 章　信息的度量

在信息论尚未作为一门学科建立起来之前，人们对于信息的度量并没有一个定量的概念，自香农开始，才将信息量的定量描述确定下来。对信息的定量描述有助于人们更方便地研究通信系统的可靠性和有效性。

在最简单的离散随机变量的情况下，下面引入关于信息度量的几个最重要的概念。

- 自信息：一个事件（消息）本身所包含的信息量，它是由事件的不确定性决定的。比如，“抛掷一枚硬币的结果是正面”这个消息包含的信息量。
- 互信息：一个事件所给出的关于另一个事件的信息量。比如，今天下雨所给出的关于明天下雨的信息量。
- 平均自信息（信息熵）：事件集（用随机变量表示）所包含的平均信息量，表示信源的平均不确定性。比如，抛掷一枚硬币的试验所包含的信息量。
- 平均互信息：一个事件集给出的关于另一个事件集的平均信息量。比如，今天的天气给出的关于明天的天气的信息量。

1.1　自信息和互信息

1.1.1　自信息

在绪论中讲过，信源发出的消息（事件）具有不确定性，而事件发生的不确定性与事件发生的概率大小有关，概率越小，不确定性越大，事件发生以后包含的信息量就越大。小概率事件的不确定性大，一旦出现必然使人感到意外，因此产生的信息量就大，特别是几乎不可能出现的事件一旦出现，必然产生极大的信息量。大概率事件是预料之中的事件，不确定性小，即使发生，也没什么信息量，特别是概率为 1 的确定事件发生以后，不会给人以任何信息量。因此，随机事件的自信息量 $I(x_i)$ 是该事件发生概率 $p(x_i)$ 的函数，并且 $I(x_i)$ 应该满足以下公理化条件：

① $I(x_i)$ 是 $p(x_i)$ 的严格递减函数。当 $p(x_1)<p(x_2)$ 时，$I(x_1)>I(x_2)$，概率越小，事件发生的不确定性越大，事件发生以后包含的自信息量越大。

② 极限情况下，当 $p(x_i)=0$ 时，$I(x_i)\to\infty$；当 $p(x_i)=1$ 时，$I(x_i)=0$。

③ 从直观概念上讲，由两个相对独立的不同消息所提供的信息量，应等于它们分别提供的信息量之和。

可以证明，满足以上公理化条件的函数形式是对数形式。

【定义 1-1】随机事件的自信息量定义为该事件发生概率的对数的负值。设事件 x_i 的概率为 $p(x_i)$，则它的自信息定义为

$$I(x_i) \overset{\text{def}}{=\!=} -\log p(x_i) = \log \frac{1}{p(x_i)} \tag{1.1}$$

自信息量的函数曲线如图 1.1 所示。可以看到，$I(x_i)$的这种定义正是满足上述公理性条件的函数形式。在它的定义域[0，1]内，自信息是非负的。

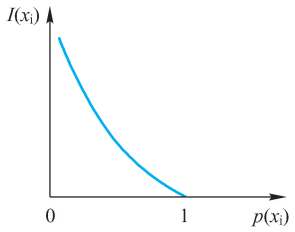


图 1.1　自信息量

$I(x_i)$代表两种含义：事件 x_i 发生前，等于事件 x_i 发生的不确定性的大小；事件 x_i 发生后，表示事件 x_i 包含或所能提供的信息量。在无噪信道中，事件 x_i 发生后，能准确无误地传输给接收者，所以 $I(x_i)$等于接收者接收到 x_i 后所获得的信息量。这是因为消除了 $I(x_i)$大小的不确定性后，才获得了这样大小的信息量。

自信息量的单位与所用对数的底有关。

通常取对数的底为 2，信息量的单位为比特（bit，binary unit）。比特是信息论中最常用的信息量单位。$p(x_i) = \frac{1}{2}$时，$I(x_i) = 1$ 比特，即概率等于$\frac{1}{2}$的事件具有 1 比特的自信息量。例如，一枚均匀硬币的任何一种抛掷结果均含有 1 比特的信息量。当取对数的底为 2 时，2 常省略。注意：计算机术语 bit 是位的单位（bit，binary digit），与信息量的单位不同，但有联系，1 位的二进制数字最大能提供 1 比特的信息量。

若取自然对数（对数以 e 为底），则自信息量的单位为奈特（nat，natural unit）。理论推导中或用于连续信源时，用以 e 为底的对数比较方便。

$$1 \text{ 奈特} = \log_2 e \text{ 比特} = 1.443 \text{ 比特}$$

工程上用以 10 为底的对数较为方便。若以 10 为对数的底，则自信息量的单位为哈特莱（Hartley）（因为哈特莱最先提出用对数来度量信息）。

$$1 \text{ 哈特莱} = \log_2 10 \text{ 比特} = 3.322 \text{ 比特}$$

如果取以 r 为底的对数($r > 1$)，则 $I(x_i) = -\log_r p(x_i)$（r 进制单位）。

$$1 \quad r \text{ 进制单位} = \log_2 r \text{ 比特}$$

【例 1.1】

（1）英文字母中“a”出现的概率为 0.064，“c”出现的概率为 0.022，分别计算它们的自信息量。

（2）假定前后字母出现是互相独立的，计算“ac”的自信息量。

（3）假定前后字母出现不是互相独立的，当“a”出现以后，“c”出现的概率为 0.04，计算“a”出现以后，“c”出现的自信息量。

【解】

（1）
$$I(a) = -\log 0.064 = 3.96 \text{ 比特}$$
$$I(c) = -\log 0.022 = 5.51 \text{ 比特}$$

（2）由于前后字母出现是互相独立的，“ac”出现的概率为 0.064×0.022。

$$\begin{aligned} I(ac) &= -\log(0.064 \times 0.022) \\ &= -(\log 0.064 + \log 0.022) \\ &= I(a) + I(c) = 9.47 \text{ 比特} \end{aligned}$$

即两个相对独立的事件的自信息量满足可加性。也就是说，由两个相对独立的事件的积事件所提供的信息量，应等于它们分别提供的信息量之和。

（3）“a”出现的条件下，“c”出现的条件概率变大，它的不确定性变小。

$$I(c \mid a) = -\log 0.04 = 4.64 \text{ 比特}$$

1.1.2 互信息

一个事件 y_j 所给出的关于另一个事件 x_i 的信息定义为互信息（Mutual Information），用 $I(x_i;y_j)$ 表示。互信息 $I(x_i;y_j)$ 是已知事件 y_j 后所消除的关于事件 x_i 的不确定性，等于事件 x_i 本身的不确定性 $I(x_i)$ 减去已知事件 y_j 后对 x_i 仍然存在的不确定性 $I(x_i \mid y_i)$。

【定义 1-2】

$$I(x_i;y_j) \overset{\text{def}}{=\!=} I(x_i) - I(x_i \mid y_j) = \log \frac{p(x_i \mid y_j)}{p(x_i)} \tag{1.2}$$

【例 1.2】某地二月份天气出现的概率分别为：晴$\frac{1}{2}$，阴$\frac{1}{4}$，雨$\frac{1}{8}$，雪$\frac{1}{8}$。某天，有人告诉你：“今天不是晴天”。把这句话作为收到的消息 y，求当收到 y 后得到的关于各种天气的信息量。

【解】把各种天气记为 x_1（晴）、x_2（阴）、x_3（雨）、x_4（雪）。当收到消息 y 后，各种天气发生的概率变成了后验概率：

$$p(x_1 \mid y) = \frac{p(x_1\, y)}{p(y)} = 0$$

$$p(x_2 \mid y) = \frac{p(x_2\, y)}{p(y)} = \frac{\frac{1}{4}}{\frac{1}{4}+\frac{1}{8}+\frac{1}{8}} = \frac{1}{2}$$

$$p(x_3 \mid y) = \frac{p(x_3\, y)}{p(y)} = \frac{\frac{1}{8}}{\frac{1}{4}+\frac{1}{8}+\frac{1}{8}} = \frac{1}{4}$$

同理，$p(x_4 \mid y) = \frac{1}{4}$。

根据互信息量的定义，可计算出 y 与各种天气之间的互信息，即得到的信息量如下：

$$I(x_1;y) = \log \frac{p(x_1 \mid y)}{p(x_1)} = \infty$$

$$I(x_2;y) = \log \frac{p(x_2 \mid y)}{p(x_2)} = \log \frac{\frac{1}{2}}{\frac{1}{4}} = 1 \text{ 比特}$$

$$I(x_3;y)=\log\frac{p(x_3\mid y)}{p(x_3)}$$

$$=\log\frac{\frac{1}{4}}{\frac{1}{8}}=1\ \text{比特}$$

$$I(x_4;y)=\log\frac{p(x_4\mid y)}{p(x_4)}$$

$$=\log\frac{\frac{1}{4}}{\frac{1}{8}}=1\ \text{比特}$$

互信息的引出使信息的传输得到了定量的表示，是信息论发展的一个重要里程碑。

1.2 平均自信息

1.2.1 平均自信息（信息熵）的概念

自信息量是信源发出某一具体消息所含有的信息量，发出的消息不同，包含的信息量也不同，所以自信息量本身为随机变量。因此，自信息量不能用来表征整个信源的不确定度。我们定义平均自信息来表征整个信源的不确定度。平均自信息量又称为信息熵、信源熵，简称熵。

因为信源具有不确定性，所以用随机变量来表示信源，用随机变量的概率分布来描述信源的不确定性。通常，一个随机变量的所有可能取值和这些取值对应的概率$[X,P(X)]$被称为它的概率空间。

假设随机变量X有q个可能的取值$x_i(i=1,2,\cdots,q)$，各种取值出现的概率为$p(x_i)$$(i=1,2,\cdots,q)$，它的概率空间表示为

$$\begin{bmatrix}X\\P(X)\end{bmatrix}=\begin{bmatrix}X=x_1 & \cdots & X=x_i & \cdots & X=x_q\\p(x_1) & \cdots & p(x_i) & \cdots & p(x_q)\end{bmatrix}$$

注意$p(x_i)$满足概率空间的基本特性：非负性，$0\leqslant p(x_i)\leqslant 1$；完备性，$\sum_{i=1}^{q}p(x_i)=1$。

【定义 1-3】随机变量X的每个可能取值的自信息$I(x_i)$的统计平均值，定义为随机变量X的平均自信息量：

$$H(X)=E[I(x_i)]=-\sum_{i=1}^{q}p(x_i)\log p(x_i)\qquad(1.3)$$

式中，q为X的所有可能取值的个数。

熵的单位也与所取的对数的底有关，根据所取的对数底的不同，可以是比特/符号、奈特/符号、哈特莱/符号，或是r进制单位/符号。通常用比特/符号为单位。

名词“熵”是香农从物理学中的热熵这一概念借用过来的，热熵是表示分子混乱程度的一个物理量，因此香农用熵来描述信源的平均不确定性。但在热力学中任何孤立系统的演

化，热熵只能增加不能减少；而在信息论中，信息熵正好相反，只会减少而不会增加。所以有人称信息熵为负熵。

信息熵 $H(X)$ 是信源的平均不确定性的描述。第 4 章中的无失真信源编码定理和它的逆定理会进一步证明，要对信源输出的消息进行无失真的编码，平均每个信源符号至少需要用 $H(X)$ 个码符号。

一般情况下，信息熵并不等于接收者平均获得的信息量。只有在无噪情况下，接收者才能准确无误地接收到信源所发出的消息，全部消除了 $H(X)$ 大小的平均不确定性，获得的平均信息量就等于 $H(X)$。而在一般情况下，接收者不能全部消除信源的平均不确定性，获得的信息量将小于信息熵。

【例 1.3】有 8 匹马参加的一场赛马比赛，获胜概率分别为 $\left\{\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{16},\frac{1}{64},\frac{1}{64},\frac{1}{64},\frac{1}{64}\right\}$。求该场赛马的熵。

【解】

$$\begin{aligned} H(X) &= -\frac{1}{2}\log\frac{1}{2}-\frac{1}{4}\log\frac{1}{4}-\frac{1}{8}\log\frac{1}{8}-\frac{1}{16}\log\frac{1}{16}-4\,\frac{1}{64}\log\frac{1}{64} \\ &= 2 \text{ 比特/符号} \end{aligned}$$

1.2.2 熵函数的性质

信息熵 $H(X)$ 是随机变量 X 的概率分布的函数，又称为熵函数。如果把概率分布 $p(x_i)$ $(i=1,2,\cdots,q)$ 记为 $p_1,p_2,\cdots,p_q$，则熵函数又可以写成概率矢量 $\boldsymbol{p}=(p_1,p_2,\cdots,p_q)$ 函数的形式，记为 $H(\boldsymbol{p})$：

$$H(X) = -\sum_{i=1}^{q} p_i \log p_i = H(p_1,p_2,\cdots,p_q) = H(\boldsymbol{p}) \tag{1.4}$$

因为概率空间的完备性，即 $\sum_{i=1}^{q} p_i = 1$，所以 $H(\boldsymbol{p})$ 是 $q-1$ 元函数。当 $q=2$ 时，因为 $p_1+p_2=1$，令其中一个概率为 p，则另一个概率为 $1-p$，熵函数可以写成 $H(p)$。

熵函数 $H(\boldsymbol{p})$ 具有以下性质。

（1）对称性

$$H(p_1,p_2,\cdots,p_q)=H(p_2,p_1,\cdots,p_q)=\cdots=H(p_q,p_1,\cdots,p_{q-1}) \tag{1.5}$$

也就是说，概率矢量 $\boldsymbol{p}=(p_1,p_2,\cdots,p_q)$ 的各分量的次序可以任意变更，熵值不变。对称性说明熵函数仅与信源的总体统计特性有关。

例如，三个信源

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{bmatrix} x_1(\text{红}) & x_2(\text{黄}) & x_3(\text{蓝}) \\ \frac{1}{3} & \frac{1}{6} & \frac{1}{2} \end{bmatrix}$$

$$\begin{bmatrix} Y \\ P(Y) \end{bmatrix} = \begin{bmatrix} y_1(\text{红}) & y_2(\text{黄}) & y_3(\text{蓝}) \\ \frac{1}{6} & \frac{1}{2} & \frac{1}{3} \end{bmatrix}$$

$$\begin{bmatrix} Z \\ P(Z) \end{bmatrix} = \begin{bmatrix} z_1(\text{晴}) & z_2(\text{雾}) & z_3(\text{雨}) \\ \frac{1}{3} & \frac{1}{6} & \frac{1}{2} \end{bmatrix}$$

的信息熵都相等，因为三个信源的总体统计特性都相同，香农熵只抽取了信源信息输出的统计特征，而没有考虑信息的具体含义和效用。

（2）确定性

$$H(1,0)=H(1,0,0)=H(1,0,0,0)=\cdots=H(1,0,\cdots,0)=0 \tag{1.6}$$

在概率矢量 $\boldsymbol{p}=(p_1,p_2,\cdots,p_q)$ 中，只要有一个分量为 1，其他分量必为 0，它们对熵的贡献均为 0，因此熵等于 0。也就是说，确定信源的平均不确定度为 0。

（3）非负性

$$H(\boldsymbol{p})=H(p_1,p_2,\cdots,p_q)\geqslant 0 \tag{1.7}$$

对确定信源，上式中的等号成立。信源熵是自信息的数学期望，自信息是非负值，所以信源熵必定是非负的。离散信源熵才有这种非负性，以后会讲到连续信源的相对熵可能出现负值。

（4）扩展性

$$\lim_{\varepsilon\to 0} H_{q+1}(p_1,p_2,\cdots,p_q-\varepsilon,\varepsilon)=H_q(p_1,p_2,\cdots,p_q) \tag{1.8}$$

这是因为 $\lim_{\varepsilon\to 0}(\varepsilon\log\varepsilon)=0$。

扩展性的含义是，增加一个基本不会出现的小概率事件，信源的熵保持不变。虽然小概率事件出现给予接收者的信息量很大，但在熵的计算中，它占的比重很小，可以忽略不计，这也是熵的总体平均性的体现。

（5）连续性

$$\lim_{\varepsilon\to 0} H(p_1,p_2,\cdots,p_{q-1}-\varepsilon,p_q+\varepsilon)=H(p_1,p_2,\cdots,p_q) \tag{1.9}$$

即信源概率空间中概率分量的微小波动，不会引起熵的变化。

（6）递增性（递推性）

$$H(p_1,p_2,\cdots,p_{n-1},q_1,q_2,\cdots,q_m)=H(p_1,p_2,\cdots,p_n)+p_nH\left(\frac{q_1}{p_n},\frac{q_2}{p_n},\cdots,\frac{q_m}{p_n}\right) \tag{1.10}$$

递增性表明，假如有一信源的 n 个元素的概率分布为 $(p_1,p_2,\cdots,p_n)$，其中某个元素 x_n 又被划分成 m 个元素，这 m 个元素的概率之和等于元素 x_n 的概率，这样得到的新信源的熵增加了一项，原因是划分产生了不确定性。

【例 1.4】利用递增性计算 $H\left(\frac{1}{2},\frac{1}{8},\frac{1}{8},\frac{1}{8},\frac{1}{8}\right)$。

【解】

$$H\left(\frac{1}{2},\frac{1}{8},\frac{1}{8},\frac{1}{8},\frac{1}{8}\right)=H\left(\frac{1}{2},\frac{1}{2}\right)+\frac{1}{2}\times H\left(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}\right)$$

$$=1+\frac{1}{2}\times 2=2 \text{ 比特/符号}$$

（7）极值性

$$H(p_1,p_2,\cdots,p_n)\leqslant H\left(\frac{1}{n},\frac{1}{n},\cdots,\frac{1}{n}\right)=\log n \tag{1.11}$$

式中，n 是随机变量 X 的可能取值的个数。

极值性表明离散信源中各消息等概率出现时熵最大，这就是最大离散熵定理。连续信源的最大熵则与约束条件有关。极值性可视为

$$H(p_1,p_2,\cdots,p_n) \leqslant -\sum_{i=1}^{n} p_i \log q_i \tag{1.12}$$

的特例情况。式（1.12）也称为 Gibbs 不等式。

【证明】利用詹森不等式（参见本章扩展阅读）：

$$\begin{aligned} H(p_1,p_2,\cdots,p_n) + \sum_{i=1}^{n} p_i \log q_i &= -\sum_{i=1}^{n} p_i \log p_i + \sum_{i=1}^{n} p_i \log q_i \\ &= \sum_{i=1}^{n} p_i \log \frac{q_i}{p_i} \\ &\leqslant \log \sum_{i=1}^{n} \left(p_i \cdot \frac{q_i}{p_i} \right) = 0 \end{aligned}$$

当 $\frac{q_i}{p_i}=1(i=1,2,\cdots,n)$ 时，等号成立。证毕。

式（1.12）表明，任一随机变量的概率分布 p_i，对其他概率分布 q_i 的自信息 $\log q_i$ 的数学期望，必不小于概率分布 p_i 本身定义的熵 $H(p_1,p_2,\cdots,p_n)$。

取 $q_i=\frac{1}{n}(i=1,2,\cdots,n)$，由式(1.12)可得到

$$H(p_1,p_2,\cdots,p_n) \leqslant H\left(\frac{1}{n},\frac{1}{n},\cdots,\frac{1}{n}\right) = \log n \tag{1.13}$$

当 $p_i=\frac{1}{n}(i=1,2,\cdots,n)$ 时，上式等号成立。

二元信源的熵函数如图 1.2 所示，当信源输出的消息等概分布时，信源熵达到最大值 1 比特/符号。因此当二元数字是由等概的二元信源输出时，每个二元数字提供 1 比特的信息量。否则，每个二元数字提供的信息量小于 1 比特。这就是信息量的单位比特和计算机术语位的单位比特的关系。

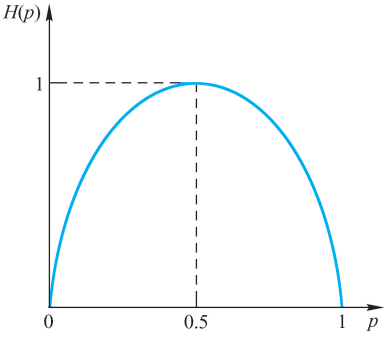


图 1.2　二元熵函数

（8）上凸性

$H(\boldsymbol{p})$是严格的上凸函数，设 $\boldsymbol{p}=(p_1,p_2,\cdots,p_q)$，$\boldsymbol{p}'=(p_1',p_2',\cdots,p_q')$，$\sum_{i=1}^{q} p_i = 1$，$\sum_{i=1}^{q} p_i' = 1$，

则对于任意小于 1 的正数 $\alpha(0<\alpha<1)$，以下不等式成立：

$$H[\alpha \boldsymbol{p}+(1-\alpha)\boldsymbol{p}']>\alpha H(\boldsymbol{p})+(1-\alpha)H(\boldsymbol{p}') \tag{1.14}$$

【证明】

因为 $0\leqslant p_i\leqslant 1$，$0\leqslant p_i'\leqslant 1$，且 $0<\alpha<1$，所以 $0\leqslant \alpha p_i+(1-\alpha)p_i'\leqslant 1$，且 $\sum_{i=1}^{q}[\alpha p_i+(1-\alpha)p_i']=1$，从而 $\alpha p_i+(1-\alpha)p_i'$ 可以视为一种新的概率分布。

$$\begin{aligned}
H[\alpha \boldsymbol{p}+(1-\alpha)\boldsymbol{p}'] &= -\sum_{i=1}^{q}[\alpha p_i+(1-\alpha)p_i']\log[\alpha p_i+(1-\alpha)p_i'] \\
&= -\alpha\sum_{i=1}^{q}p_i\log[\alpha p_i+(1-\alpha)p_i']-(1-\alpha)\sum_{i=1}^{q}p_i'\log[\alpha p_i+(1-\alpha)p_i'] \\
&\geqslant -\alpha\sum_{i=1}^{q}p_i\log p_i-(1-\alpha)\sum_{i=1}^{q}p_i'\log p_i' \\
&\geqslant \alpha H(\boldsymbol{p})+(1-\alpha)H(\boldsymbol{p}')
\end{aligned}$$

如果 $\boldsymbol{p}\neq\boldsymbol{p}'$，则

$$H[\alpha \boldsymbol{p}+(1-\alpha)\boldsymbol{p}']>\alpha H(\boldsymbol{p})+(1-\alpha)H(\boldsymbol{p}') \tag{1.15}$$

成立。

证毕。

上凸函数在定义域内的极值必为极大值，利用熵函数的这个性质可以证明熵函数的极值性。请读者自证。

直观来看，随机变量的不确定程度并不都是一样的。例如，抛掷一枚均匀硬币时，结果所得到的信息量会比抛掷一枚偏畸硬币时所得到的信息量大，投掷一颗均匀骰子的试验比抛掷一枚均匀硬币的试验所得到的信息量大。怎么度量这种不确定性呢？香农指出，存在这样的不确定性的度量，它是随机变量的概率分布的函数，而且必须满足如下 3 个公理性条件：

① 连续性条件：$f(p_1,p_2,\cdots,p_n)$ 应是 $p_i(i=1,2,\cdots,n)$ 的连续函数。

② 等概时为单调函数：$f\left(\frac{1}{n},\frac{1}{n},\cdots,\frac{1}{n}\right)$ 应是 n 的增函数。

③ 递增性条件：当随机变量的取值不是通过一次试验而是通过若干次试验才最后得到时，随机变量在每次试验中的不确定性应该可加，且其和始终与通过一次试验取得的不确定程度相同，即

$$\begin{aligned}
f(p_1,p_2,\cdots,p_n)=&f[(p_1+p_2+\cdots+p_k),p_{k+1},\cdots,p_n]+\\
&(p_1+p_2+\cdots+p_k)f(p_1',p_2',\cdots,p_k')
\end{aligned}$$

其中，$p_k'=\dfrac{p_k}{(p_1+p_2+\cdots+p_k)}$。

香农根据这 3 个公理性条件于 1948 年首先提出了“熵”的概念，并没有像现在这样把熵视为自信息的均值。后来，范恩斯坦（Feinstein）等人从数学上严格地证明了当满足上述条件时，信息熵的表达形式是唯一的。

1.2.3 联合熵与条件熵

一个随机变量的不确定性可以用熵来表示，这个概念可以推广到多个随机变量。

【定义 1-4】二维随机变量 XY 的概率空间表示为

$$\begin{bmatrix} XY \\ P(XY) \end{bmatrix} = \begin{bmatrix} x_1y_1 & \cdots & x_iy_j & \cdots & x_ny_m \\ p(x_1y_1) & \cdots & p(x_iy_j) & \cdots & p(x_ny_m) \end{bmatrix}$$

其中，$p(x_iy_j)$ 满足概率空间的非负性和完备性：$0 \leqslant p(x_iy_j) \leqslant 1$，$\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_iy_j) = 1$。

二维随机变量 XY 的联合熵定义为联合自信息的数学期望，它是二维随机变量 XY 的不确定性的度量：

$$H(XY) \overset{\text{def}}{=\!=} \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_iy_j)I(x_iy_j) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_iy_j)\log p(x_iy_j) \tag{1.16}$$

给定 $X = x_i$ 的条件下，随机变量 Y 的不确定性为

$$H(Y \mid x_i) = -\sum_{j} p(y_j \mid x_i)\log p(y_j \mid x_i) \tag{1.17}$$

不同的 x_i 对应的 $H(Y \mid x_i)$ 是变化的。对 $H(Y \mid x_i)$ 的所有可能值进行统计平均，就可以得出给定 X 时 Y 的条件熵 $H(Y \mid X)$。

【定义 1-5】

$$\begin{aligned} H(Y \mid X) &= \sum_{i} p(x_i)H(Y \mid x_i) \\ &= -\sum_{i}\sum_{j} p(x_i)p(y_j \mid x_i)\log p(y_j \mid x_i) \\ &= -\sum_{i}\sum_{j} p(x_iy_j)\log p(y_j \mid x_i) \end{aligned} \tag{1.18}$$

$H(Y \mid X)$ 表示已知 X 时 Y 的不确定性。同理，有

$$H(X \mid Y) = -\sum_{i}\sum_{j} p(x_iy_j)\log p(x_i \mid y_j) \tag{1.19}$$

各类熵之间的关系如下。

（1）联合熵与信息熵、条件熵的关系

$$H(XY) = H(X) + H(Y \mid X) \tag{1.20}$$

【证明】

$$\begin{aligned} H(XY) &= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_iy_j)\log p(x_iy_j) \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_iy_j)\log[p(x_i)p(y_j \mid x_i)] \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_iy_j)\log p(x_i) - \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_iy_j)\log p(y_j \mid x_i) \\ &= -\sum_{i=1}^{n}\left[\sum_{j=1}^{m} p(x_iy_j)\right]\log p(x_i) - \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i)p(y_j \mid x_i)\log p(y_j \mid x_i) \\ &= -\sum_{i=1}^{n} p(x_i)\log p(x_i) - \sum_{i=1}^{n} p(x_i)\sum_{j=1}^{m} p(y_j \mid x_i)\log p(y_j \mid x_i) \\ &= H(X) + \sum_{i=1}^{n} p(x_i)H(Y \mid x_i) \\ &= H(X) + H(Y \mid X) \end{aligned}$$

上述证明还可以更简洁地表示为

$$
\begin{aligned}
H(XY) &= E\left[\log\frac{1}{p(xy)}\right] \\
&= E\left[\log\frac{1}{p(x)p(y|x)}\right] \\
&= E\left[\log\frac{1}{p(x)}+\log\frac{1}{p(y|x)}\right] \\
&= E\left[\log\frac{1}{p(x)}\right]+E\left[\log\frac{1}{p(y|x)}\right] \\
&= H(X)+H(Y|X)
\end{aligned}
$$

即两个随机变量 X 和 Y 的联合熵等于 X 的熵加上在 X 已知条件下 Y 的条件熵。这个关系可以方便地推广到 N 个随机变量的情况：

$$H(X_1X_2\cdots X_N)=H(X_1)+H(X_2|X_1)+\cdots+H(X_N|X_1X_2\cdots X_{N-1}) \tag{1.21}$$

这称为熵函数的链规则。

证毕。

【推论】 当二维随机变量 X、Y 相互独立时，联合熵等于 X 和 Y 各自熵之和：

$$H(XY)=H(X)+H(Y) \tag{1.22}$$

【证明】

因为随机变量 X 与 Y 相互独立，所以

$$
\begin{aligned}
p(x_iy_j) &= p(x_i)p(y_j) \\
H(XY) &= E[-\log p(xy)] \\
&= E[-\log p(x)p(y)] \\
&= E[-(\log p(x)+\log p(y))] \\
&= E[-\log p(x)]+E[-\log p(y)] \\
&= H(X)+H(Y)
\end{aligned}
$$

如果 N 个随机变量 $X_1,X_2,\cdots,X_N$ 相互独立，则

$$H(X_1X_2\cdots X_N)=\sum_{i=1}^{N}H(X_i) \tag{1.23}$$

（2）条件熵与信息熵的关系

$$H(X|Y)\leqslant H(X) \tag{1.24}$$

$$H(Y|X)\leqslant H(Y) \tag{1.25}$$

【证明】

利用式(1.12)证明式(1.24)。

$$
\begin{aligned}
H(X|Y)-\sum_i\sum_j p(x_iy_j)\log p(x_i|y_j) &= -\sum_i\sum_j p(y_j)p(x_i|y_j)\log p(x_i|y_j) \\
&= -\sum_j p(y_j)\sum_i p(x_i|y_j)\log p(x_i|y_j) \\
&\leqslant -\sum_j p(y_j)\sum_i p(x_i|y_j)\log p(x_i) \\
&= -\sum_i\sum_j p(x_iy_j)\log p(x_i) \\
&= -\sum_i p(x_i)\log p(x_i) \\
&= H(X)
\end{aligned}
$$

$p(x_i \mid y_j) = p(x_i)$时，上式中的等号成立。

类似地可以证明 $H(Y \mid X) \leqslant H(Y)$。

证毕。

（3）联合熵和信息熵的关系

$$H(XY) \leqslant H(X) + H(Y) \tag{1.26}$$

当 X、Y 相互独立时，等号成立。

【证明】 $H(XY) = H(X) + H(Y \mid X) \leqslant H(X) + H(Y)$

当 X、Y 相互独立时，等号成立。

推广到 N 个随机变量的情况：

$$H(X_1X_2\cdots X_N) \leqslant H(X_1) + H(X_2) + \cdots + H(X_N) \tag{1.27}$$

当 $X_1, X_2, \cdots, X_N$ 相互独立时，等号成立。

【例 1.5】随机变量 X、Y 的联合概率分布见表 1.1，求联合熵 $H(XY)$ 和条件熵 $H(Y \mid X)$。

表 1.1 联合概率分布

X \ Y	0	1
0	$\frac{1}{4}$	$\frac{1}{4}$
1	$\frac{1}{2}$	0

【解】

$$H(XY) = \frac{1}{4}\log\frac{1}{\frac{1}{4}} + \frac{1}{4}\log\frac{1}{\frac{1}{4}} + \frac{1}{2}\log\frac{1}{\frac{1}{2}}$$

$$= \frac{2}{4}\log 4 + \frac{1}{2}\log 2$$

$$= \frac{2}{4}\cdot 2 + \frac{1}{2}\cdot 1 = \frac{3}{2}\text{比特/联合符号}$$

由联合概率分布，得到 X 的边沿概率分布如下：

$$P(X=0) = \frac{1}{2}$$

$$P(X=1) = \frac{1}{2}$$

条件概率分布见表 1.2。

表 1.2 条件概率分布

X \ Y	0	1
0	$\frac{1}{2}$	$\frac{1}{2}$
1	1	0

从而得到

$$H(Y\mid X=0)=1\ \text{比特/符号}$$
$$H(Y\mid X=1)=0\ \text{比特/符号}$$
$$H(Y\mid X)=\frac{1}{2}\times 1+\frac{1}{2}\times 0=\frac{1}{2}\text{比特/符号}$$

注意：

$$H(Y)=H\left(\frac{1}{4}\right)=0.8113>\frac{1}{2}=H(Y\mid X)$$

1.3 平均互信息

1.3.1 平均互信息的概念

互信息 $I(x_i;y_j)$ 表示事件 y_j 所给出的关于另一个事件 x_i 的信息，它随 x_i 和 y_j 的变化而变化。为了从整体上表示从一个随机变量 Y 所给出的关于另一个随机变量 X 的信息量，我们定义互信息 $I(x_i;y_j)$ 在 XY 的联合概率空间中的统计平均值为随机变量 X 和 Y 间的平均互信息。

【定义 1-6】

$$\begin{aligned}I(X;Y)&=\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_iy_j)I(x_i;y_j)\\&=\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_iy_j)\log\frac{p(x_i\mid y_j)}{p(x_i)}\qquad(1.28)\\&=\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_iy_j)\log\frac{1}{p(x_i)}-\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_iy_j)\log\frac{1}{p(x_i\mid y_j)}\\&=H(X)-H(X\mid Y)\qquad(1.29)\end{aligned}$$

条件熵 $H(X\mid Y)$ 表示给定随机变量 Y 后，对随机变量 X 仍然存在的不确定度。所以 Y 关于 X 的平均互信息是收到 Y 前后关于 X 的不确定度减少的量，也就是从 Y 所获得的关于 X 的平均信息量。

【例 1.6】掷骰子。若结果是 1、2、3 或 4，则抛一次硬币；若结果是 5 或 6，则抛两次硬币。试计算从抛硬币的结果可以得到多少掷骰子的信息量。

【解】

本题的题意是根据抛硬币出现正面的次数 Y 来获得关于掷骰子结果 X 的信息（两种结果）。

设掷骰子结果是 1、2、3、4 的事件为 $X=0$，结果是 5、6 的事件为 $X=1$，随机变量 $Y=0$ 表示抛币出现 0 次正面，$Y=1$ 表示出现 1 次正面，$Y=2$ 表示出现 2 次正面。

随机变量 X 的概率空间为

$$\begin{bmatrix}X\\P\end{bmatrix}=\begin{bmatrix}0 & 1\\\frac{2}{3} & \frac{1}{3}\end{bmatrix}$$

条件概率矩阵为

$$\boldsymbol{P}_{Y|X}=\begin{bmatrix}\frac{1}{2} & \frac{1}{2} & 0\\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4}\end{bmatrix}$$

Y 的概率分布为

$$\begin{aligned}\boldsymbol{P}_Y &= \boldsymbol{P}_X\boldsymbol{P}_{Y|X}\\ &=\begin{bmatrix}\frac{2}{3} & \frac{1}{3}\end{bmatrix}\begin{bmatrix}\frac{1}{2} & \frac{1}{2} & 0\\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4}\end{bmatrix}\\ &=\begin{bmatrix}\frac{5}{12} & \frac{1}{2} & \frac{1}{12}\end{bmatrix}\end{aligned}$$

所以，Y 的信息熵为

$$\begin{aligned}H(Y) &= p_Y(0)\log\frac{1}{p_Y(0)}+p_Y(1)\log\frac{1}{p_Y(1)}+p_Y(2)\log\frac{1}{p_Y(2)}\\ &=\frac{5}{12}\log\frac{12}{5}+\frac{1}{2}\log 2+\frac{1}{12}\log 12=1.325\ \text{比特/符号}\end{aligned}$$

又可以根据 X 的概率分布和条件概率分布 $\boldsymbol{P}_{Y|X}$ 求出 $H(Y|X)$：

$$\begin{aligned}H(Y|X) &=\frac{2}{3}\left(\frac{1}{2}\log 2+\frac{1}{2}\log 2\right)+\frac{1}{3}\left(\frac{1}{4}\log 4+\frac{1}{2}\log 2+\frac{1}{4}\log 4\right)\\ &=1.166\ \text{比特/符号}\end{aligned}$$

$$I(X;Y)=H(Y)-H(Y|X)=1.325-1.166=0.159\ \text{比特/符号}$$

即从抛硬币出现正面次数平均得到关于掷骰子结果的信息量为 0.159 比特/符号。

1.3.2 平均互信息的性质

（1）非负性

$$I(X;Y)\geqslant 0 \tag{1.30}$$

【证明】

$$\begin{aligned}-I(X;Y) &= \sum_{i=1}^{n}\sum_{j=1}^{m}p(x_iy_j)\log\frac{p(x_i)p(y_j)}{p(x_iy_j)}\\ &\leqslant \log\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_iy_j)\frac{p(x_i)p(y_j)}{p(x_iy_j)}\\ &= \log\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_i)p(y_j)\\ &= 0\end{aligned}$$

所以，$I(X;Y)\geqslant 0$。

证毕。

平均互信息是非负的，说明给定随机变量 Y 后，一般总能消除一部分关于 X 的不确定性。

（2）互易性（对称性）

$$I(X;Y)=I(Y;X) \tag{1.31}$$

【证明】

$$
\begin{aligned}
I(X;Y) &= \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i y_j)\log\frac{p(x_i \mid y_j)}{p(x_i)} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i y_j)\log\frac{p(x_i y_j)}{p(x_i)p(y_j)} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i y_j)\log\frac{p(y_j \mid x_i)}{p(y_j)} \\
&= I(Y;X)
\end{aligned}
$$

证毕。

对称性表示从 Y 中获得关于 X 的信息量等于从 X 中获得关于 Y 的信息量。

（3）平均互信息与各类熵的关系（如图 1.3 所示）

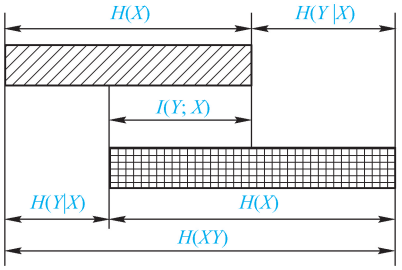


图 1.3　平均互信息与各类熵之间的关系

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) \\
&= H(X) + H(Y) - H(XY)
\end{aligned}
\tag{1.32}
$$

当 X、Y 统计独立时，$I(X;Y)=0$。

（4）极值性

$$
I(X;Y) \leqslant H(X), \quad I(X;Y) \leqslant H(Y) \tag{1.33}
$$

由于 $I(X;Y)=H(X)-H(X \mid Y)=H(Y)-H(Y \mid X)$，而条件熵 $H(X \mid Y)$、$H(Y \mid X)$ 是非负的，所以可得到

$$
I(X;Y) \leqslant H(X), \ I(X;Y) \leqslant H(Y)
$$

极值性说明，从一个事件提取关于另一个事件的信息量，至多只能是另一个事件的平均自信息量那么多，不会超过另一个事件本身所含的信息量。最好的情况是，通信后 $I(X;Y)=H(X)=H(Y)$。最坏的情况是，当 X、Y 相互独立时，从一个事件不能得到另一个事件的任何信息，即 $I(X;Y)=0$，等效于通信中断。

（5）凸函数性

【定理 1-1】 当条件概率分布 $\{p(y_j \mid x_i)\}$ 给定时，平均互信息 $I(X;Y)$ 是输入分布 $\{p(x_i)\}$ 的上凸函数。

【证明】

设固定条件概率分布为 $\{p(y_j \mid x_i)\}$。

$\{p_1(x_i)\}$ 和 $\{p_2(x_i)\}$ 为信源的两种不同的概率分布，相应的平均互信息记为 $I[p_1(x_i)]$ 和 $I[p_2(x_i)]$，再选择信源符号集的另一种概率分布 $\{p(x_i)\}$，且令

$$p(x_i)=\theta p_1(x_i)+(1-\theta)p_2(x_i) \tag{1.34}$$

其中，$0<\theta<1$，相应的平均互信息记为 $I[p(x_i)]$。

根据上凸函数的定义，我们需要证明

$$I[p(x_i)]\geqslant\theta I[p_1(x_i)]+(1-\theta)I[p_2(x_i)] \tag{1.35}$$

根据平均互信息的定义，有

$$\begin{aligned}
&\theta I[p_1(x_i)]+(1-\theta)I[p_2(x_i)]-I[p(x_i)]\\
=&\ \theta\sum_{i,j}p_1(x_i)p(y_j\mid x_i)\log\frac{p(y_j\mid x_i)}{p_1(y_j)}+(1-\theta)\sum_{i,j}p_2(x_i)p(y_j\mid x_i)\log\frac{p(y_j\mid x_i)}{p_2(y_j)}-\\
&\sum_{i,j}p(x_i)p(y_j\mid x_i)\log\frac{p(y_j\mid x_i)}{p(y_j)}\\
=&\ \theta\sum_{i,j}p_1(x_i)p(y_j\mid x_i)\log\frac{p(y_j\mid x_i)}{p_1(y_j)}+(1-\theta)\sum_{i,j}p_2(x_i)p(y_j\mid x_i)\log\frac{p(y_j\mid x_i)}{p_2(y_j)}-\\
&\theta\sum_{i,j}p_1(x_i)p(y_j\mid x_i)\log\frac{p(y_j\mid x_i)}{p(y_j)}-(1-\theta)\sum_{i,j}p_2(x_i)p(y_j\mid x_i)\log\frac{p(y_j\mid x_i)}{p(y_j)}\\
&\qquad\qquad\qquad

$\{p_1(y_j \mid x_i)\}$ 和 $\{p_2(y_j \mid x_i)\}$ 为信道的两种不同的转移概率，相应的平均互信息记为 $I[p_1(y_j \mid x_i)]$ 和 $I[p_2(y_j \mid x_i)]$，再选择信道的另一种转移概率 $\{p(y_j \mid x_i)\}$，且令

$$p(y_j \mid x_i) = \theta p_1(y_j \mid x_i) + (1-\theta)p_2(y_j \mid x_i) \tag{1.36}$$

其中，$0 < \theta < 1$，相应的平均互信息记为 $I[\theta(y_j \mid x_i)]$。

根据下凸函数的定义，我们需要证明

$$\theta I[p_1(y_j \mid x_i)] + (1-\theta) I[p_2(y_j \mid x_i)] \geqslant I[p(y_j \mid x_i)] \tag{1.37}$$

根据平均互信息的定义，有

$$\begin{aligned}
&I[p(y_j \mid x_i)] - \theta I[p_1(y_j \mid x_i)] - (1-\theta) I[p_2(y_j \mid x_i)] \\
=& \sum_{i,j} p(x_i)p(y_j \mid x_i) \log \frac{p(x_i \mid y_j)}{p(x_i)} - \theta \sum_{i,j} p(x_i)p_1(y_j \mid x_i) \log \frac{p_1(x_i \mid y_j)}{p(x_i)} - \\
&(1-\theta) \sum_{i,j} p(x_i)p_2(y_j \mid x_i) \log \frac{p_2(x_i \mid y_j)}{p(x_i)} \\
=& \theta \sum_{i,j} p(x_i)p_1(y_j \mid x_i) \log \frac{p(x_i \mid y_j)}{p(x_i)} + (1-\theta) \sum_{i,j} p(x_i)p_2(y_j \mid x_i) \log \frac{p(x_i \mid y_j)}{p(x_i)} - \\
&\theta \sum_{i,j} p(x_i)p_1(y_j \mid x_i) \log \frac{p_1(x_i \mid y_j)}{p(x_i)} - (1-\theta) \sum_{i,j} p(x_i)p_2(y_j \mid x_i) \log \frac{p_2(x_i \mid y_j)}{p(x_i)}
\end{aligned}$$

［将式(1.36)代入］

$$= \theta \sum_{i,j} p(x_i)p_1(y_j \mid x_i) \log \frac{p(x_i \mid y_j)}{p_1(x_i \mid y_j)} + (1-\theta) \sum_{i,j} p(x_i)p_2(y_j \mid x_i) \log \frac{p(x_i \mid y_j)}{p_2(x_i \mid y_j)}$$

（合并同类项）

$$\leqslant \theta \log \sum_{i,j} p(x_i)p_1(y_j \mid x_i) \frac{p(x_i \mid y_j)}{p_1(x_i \mid y_j)} + (1-\theta) \log \sum_{i,j} p(x_i)p_2(y_j \mid x_i) \frac{p(x_i \mid y_j)}{p_2(x_i \mid y_j)}$$

（Jensen 不等式）

$$\begin{aligned}
&= \theta \log \sum_j p(y_j) \sum_i p(x_i \mid y_j) + (1-\theta) \log \sum_j p(y_j) \sum_i p(x_i \mid y_j) \\
&= \theta \log 1 + (1-\theta) \log 1 \\
&= 0
\end{aligned}$$

所以证得式(1.36)。

以上证明中应用了如下概率关系

$$\begin{aligned}
p(x_i)p_1(y_j \mid x_i) &= p_1(x_i y_j) = p_1(x_i \mid y_j)p(y_j) \\
p(x_i)p_2(y_j \mid x_i) &= p_2(x_i y_j) = p_2(x_i \mid y_j)p(y_j) \\
\sum_i p(x_i \mid y_j) &= 1 \\
\sum_j p(y_j) &= 1
\end{aligned}$$

因此，由下凸函数的定义可知，在给定输入分布的情况下，平均互信息量 $I(X;Y)$ 是条件概率分布 $\{p(y_j \mid x_i)\}$ 的下凸函数。如果把条件概率分布 $\{p(y_j \mid x_i)\}$ 看成信道的转移概率分布，那么对于给定的输入分布，必定存在一种最差的信道，此信道的干扰（噪声）最大，接收者获得的信息量最小。在第 6 章中讨论信息率失真函数时会用到这个定理。

1.3.3 数据处理定理

为了研究数据处理定理，我们需要引入三元随机变量 X、Y、Z 的平均条件互信息和平均联合互信息的概念。

【定义 1-7】平均条件互信息

$$\begin{aligned} I(X;Y|Z) &= E[I(x;y|z)] \\ &= \sum_x \sum_y \sum_z p(xyz)\log\frac{p(x|yz)}{p(x|z)} \end{aligned} \tag{1.38}$$

它表示随机变量 Z 给定后，从随机变量 Y 所得到的关于随机变量 X 的信息量。

【定义 1-8】平均联合互信息

$$\begin{aligned} I(X;YZ) &= E[I(x;yz)] \\ &= \sum_x \sum_y \sum_z p(xyz)\log\frac{p(x|yz)}{p(x)} \end{aligned} \tag{1.39}$$

它表示从二维随机变量 YZ 所得到的关于随机变量 X 的信息量。

可以证明

$$\begin{aligned} I(X;YZ) &= \sum_x \sum_y \sum_z p(xyz)\log\frac{p(x|z)\cdot p(x|yz)}{p(x)\cdot p(x|z)} \\ &= I(X;Z) + I(X;Y|Z) \end{aligned} \tag{1.40}$$

同理

$$I(X;YZ) = I(X;Y) + I(X;Z|Y) \tag{1.41}$$

【定理 1-3】数据处理定理：如果随机变量 X、Y、Z 构成一个马尔可夫链，则有以下关系成立：

$$\begin{cases} I(X;Z) \leqslant I(X;Y) \\ I(X;Z) \leqslant I(Y;Z) \end{cases} \tag{1.42}$$

等号成立的条件是对于任意的 x、y、z，有 $p(x|yz)=p(x|z)$ 和 $p(z|xy)=p(z|x)$。

【证明】

当 X、Y、Z 构成一个马尔可夫链时，Y 值给定后，X、Z 可以认为是互相独立的。所以 $I(X;Z|Y)=0$

又因为 $I(X;YZ)=I(X;Y)+I(X;Z|Y)=I(X;Z)+I(X;Y|Z)$，并且 $I(X;Y|Z)\geqslant 0$，所以 $I(X;Z)\leqslant I(X;Y)$。

当 $p(x|yz)=p(x|z)$ 时，Z 值给定后，X 和 Y 相互独立，因而 $I(X;Y|Z)=0$，所以 $I(X;Z)=I(X;Y)$。这时 $p(x|yz)=p(x|z)=p(x|y)$。Y、Z 为确定关系时显然满足该条件。

同理可以证明，$I(X;Z)\leqslant I(Y;Z)$，且当 $p(z|xy)=p(z|x)$ 时，等号成立。

证毕。

$I(X;Z)\leqslant I(X;Y)$ 表明从 Z 所得到的关于 X 的信息量，小于等于从 Y 所得到的关于 X 的信息量。如果把 $Y\to Z$ 视为数据处理系统，那么通过数据处理后，虽然可以满足我们的某种具体要求，但是从信息量来看，处理后会损失一部分信息，最多保持原来获得的信息。也就是说，对接收到的数据 Y 进行处理后，决不会减少关于 X 的不确定性。这个定理称为数据处理定理。数据处理定理与我们日常生活中的经验是一致的。比如，我们知道，通过别人转

述一段话或多或少会有一些失真，通过书本得到的间接经验总不如直接经验来得准确。

数据处理定理再次说明，在任何信息传输系统中，最后获得的信息至多是信源所提供的信息，如果在某一过程中丢失了一些信息，以后的系统不管如何处理，如不触及丢失信息的输入端，就不能再恢复已丢失的信息。这就是信息不增性原理，它与热熵不减原理正好对应，反映了信息的物理意义。

1.3.4 相对熵（KL 散度）

相对熵（Relative Entropy）又称为 KL 散度（Kullback – Leibler Divergence，KLD）、KL 距离、信息散度（Information Divergence）、鉴别信息等。

设 $p(x)$ 和 $q(x)$ 是随机变量 X 的两个概率概率分布。典型情况下，$p(x)$ 表示 X 的真实分布，$q(x)$ 表示 X 的理论分布、模型分布或 $p(x)$ 的近似分布，则 $p(x)$ 对 $q(x)$ 的相对熵为

$$D(p \parallel q) = \sum_{i=1}^{n} p(x) \log \frac{p(x)}{q(x)}$$

相对熵可以用来衡量两个分布的差异，或者说距离，所以又被称为 KL 距离。当两个随机分布相同时，它们的相对熵为零，当两个随机分布的差别增大时，它们的相对熵也会增大，是两个概率分布的差别的非对称性的度量。例如，一个符号集有 4 个符号，方法 A 得到 4 个符号的概率分别是 0.1、0.2、0.3、0.4，方法 B（或者说是事实情况）得到 4 个符号的概率分别是 0.4、0.3、0.2、0.1，那么这两个分布的相对熵为

$$D(A \parallel B) = 0.1 \log\left(\frac{0.1}{0.4}\right) + 0.2 \log\left(\frac{0.2}{0.3}\right) + 0.3 \log\left(\frac{0.3}{0.2}\right) + 0.4 \log\left(\frac{0.4}{0.1}\right)$$

$$D(B \parallel A) \neq D(A \parallel B)$$

平均互信息可视为相对熵的一个特例。相对熵有两个主要性质：

① 尽管直观上相对熵是个度量或距离函数，但它并不是一个真正的度量或距离，因为它不具有对称性，即 $D(p \parallel q) \neq D(q \parallel p)$。

② 相对熵的值为非负值，即 $D(p \parallel q) \geqslant 0$。当且仅当两分布相同时，等号成立。利用 Gibbs 不等式可以证明相对熵的非负性。

利用相对熵可以比较文本的相似度，先统计出词的频率，然后计算相对熵即可。另外，相对熵可以度量使用基于概率分布 $Q(X)$ 的编码方案为真实概率分布为 $P(X)$ 的样本进行编码平均所需的额外的比特数。根据香农的信息论，给定一个信源的概率分布为 $P(X)$，我们可以设计一种编码，使得表示该信源符号平均需要的比特数最少，最少比特数等于这个信源的熵 $H(p)$。在不知道它的真实分布的情况下，假设它的概率分布为 $Q(X)$，即用概率分布 $Q(X)$ 的最优编码来为概率分布是 $P(X)$ 的字符集编码，那么表示这些字符就会比理想情况多用一些比特数，多出来的比特数就是相对熵：

$$\begin{aligned} E_p[-\log q] &= -\sum_x p(x) \log q(x) \\ &= H(p) + D(p \parallel q) \end{aligned}$$

其中，$E_p[-\log q] = -\sum_x p(x) \log q(x)$，又称为交叉熵（Cross Entropy）。

本章介绍的这些度量中，信息熵用来度量随机变量的不确定性，互信息用来度量随机变量的相似性，相对熵、交叉熵和条件熵用来度量随机变量的相异性。

扩展阅读： 凸函数及詹森不等式

【定义】对于任意小于 1 的正数 $\alpha(0<\alpha<1)$及定义域内的任意变量 x_1、$x_2(x_1\neq x_2)$，如果 $f[\alpha x_1+(1-\alpha)x_2]\geqslant \alpha f(x_1)+(1-\alpha)f(x_2)$，则称 $f(x)$为定义域上的上凸函数。若式中“$>$”成立，则称为严格的上凸函数。

如果 $f[\alpha x_1+(1-\alpha)x_2]\leqslant \alpha f(x_1)+(1-\alpha)f(x_2)$，则称 $f(x)$为定义域上的下凸函数，若式中“$<$”成立，则称为严格的下凸函数。

在上凸函数的任意两点之间画一条割线，函数总在割线上方。如果 $f(x)$是上凸函数，则 $-f(x)$是下凸函数。如果 $f(x)$存在非负的二阶导数，则为下凸函数，如图 1.4 所示。

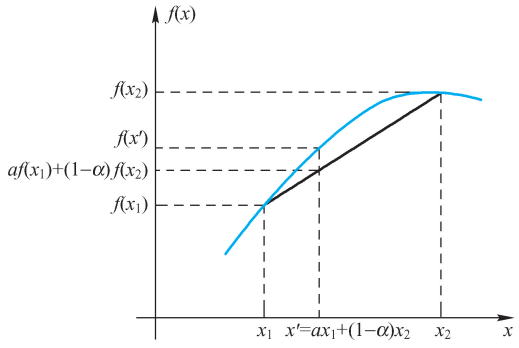


图 1.4　凸函数

对于凸函数，有一个很重要的不等式，即詹森不等式。在信息论中关于熵函数的证明经常要用到这个不等式：

若 $f(x)$是定义在区间 $[a,b]$上的实值连续上凸函数，对任意一组变量 $x_1,x_2,\cdots,x_q\in[a,b]$和任意一组非负实数 λ_1，λ_1，$\cdots$，λ_q 满足 $\sum_{k=1}^{q}\lambda_k=1$，则有

$$\sum_{k=1}^{q}\lambda_k f(x_k)\leqslant f\left[\sum_{k=1}^{q}\lambda_k x_k\right]$$

【证明】

当 $q=2$ 时，由上凸函数的定义可知上式成立，因此我们用数学归纳法。

当 $q=2$ 时，上式成立，即 $\lambda_1 f(x_1)+\lambda_2 f(x_2)\leqslant f(\lambda_1 x_1+\lambda_2 x_2)$，$\lambda_1$、$\lambda_2$ 为满足 $\lambda_1+\lambda_2=1$ 的任意非负实数。

当 $q=3$ 时，则

$$\begin{aligned}
&\lambda_1 f(x_1)+\lambda_2 f(x_2)+\lambda_3 f(x_3)\\
&=(\lambda_1+\lambda_2)\left[\frac{\lambda_1}{\lambda_1+\lambda_2}f(x_1)+\frac{\lambda_2}{\lambda_1+\lambda_2}f(x_2)\right]+\lambda_3 f(x_3)\\
&\leqslant(\lambda_1+\lambda_2)f\left[\frac{\lambda_1}{\lambda_1+\lambda_2}x_1+\frac{\lambda_2}{\lambda_1+\lambda_2}x_2\right]+\lambda_3 f(x_3)\\
&\leqslant f\left[(\lambda_1+\lambda_2)\left(\frac{\lambda_1}{\lambda_1+\lambda_2}x_1+\frac{\lambda_2}{\lambda_1+\lambda_2}x_2\right)+\lambda_3 x_3\right]
\end{aligned}$$

$$=f(\lambda_1 x_1+\lambda_2 x_2+\lambda_3 x_3)$$

假设当 $q=n$ 时成立，即 $\sum_{k=1}^{n}\lambda_k f(x_k)\leqslant f\left[\sum_{k=1}^{n}\lambda_k x_k\right]$，那么当 $q=n+1$ 时，令 $\alpha=\sum_{k=1}^{n}\lambda_k$，$\lambda_{n+1}=1-\alpha$，则

$$\begin{aligned}
&\lambda_1 f(x_1)+\lambda_2 f(x_2)+\cdots+\lambda_n f(x_n)+\lambda_{n+1}f(x_{n+1})\\
&=\alpha\left[\frac{\lambda_1}{\alpha}f(x_1)+\frac{\lambda_2}{\alpha}f(x_2)+\cdots+\frac{\lambda_n}{\alpha}f(x_n)\right]+\lambda_{n+1}f(x_{n+1})\\
&=\alpha\left[\frac{\lambda_1}{\alpha}f(x_1)+\frac{\lambda_2}{\alpha}f(x_2)+\cdots+\frac{\lambda_n}{\alpha}f(x_n)\right]+(1-\alpha)f(x_{n+1})\\
&\leqslant \alpha f\left[\sum_{k=1}^{n}\frac{\lambda_k}{\alpha}x_k\right]+(1-\alpha)f(x_{n+1})\\
&\leqslant f\left[\sum_{k=1}^{n}\lambda_k x_k+\lambda_{n+1}x_{n+1}\right]\\
&=f\left[\sum_{k=1}^{n+1}\lambda_k x_k\right]
\end{aligned}$$

所以对于任意一组变量 $x_1,x_2,\cdots,x_q\in[a,b]$ 和任意一组满足 $\sum_{k=1}^{q}\lambda_k=1$ 的非负实数 $\lambda_1,\lambda_1,\cdots,\lambda_q$，有 $\sum_{k=1}^{q}\lambda_k f(x_k)\leqslant f\left[\sum_{k=1}^{q}\lambda_k x_k\right]$ 成立。

证毕。

当 $x_1,x_2,\cdots,x_q$ 视为随机变量 X 的可能取值，而 $\lambda_1,\lambda_1,\cdots,\lambda_q$ 视为对应的概率时，上式可记为 $E[f(X)]\leqslant f[E(X)]$，即函数的均值≤均值的函数。

对数函数即为上凸函数，这时上式可表示为

$$E(\log X)\leqslant\log[E(X)]$$

对于下凸函数有 $E[f(X)]\geqslant f[E(X)]$，即函数的均值≥均值的函数。

詹森不等式可以推广到多维随机变量的情况：若 $f(\boldsymbol{x})=f(x_1,x_2,\cdots,x_n)$ 为一多维函数，若对于任意小于 1 的正数 $\alpha(0<\alpha<1)$ 以及函数 $f(\boldsymbol{x})$ 定义域内的任意矢量 $\boldsymbol{x}_1$、$\boldsymbol{x}_2(\boldsymbol{x}_1\neq\boldsymbol{x}_2)$，如果 $f[\alpha\boldsymbol{x}_1+(1-\alpha)\boldsymbol{x}_2]\geqslant\alpha f(\boldsymbol{x}_1)+(1-\alpha)f(\boldsymbol{x}_2)$，则称 $f(\boldsymbol{x})$ 为定义域上的上凸函数。若式中“>”成立，则称为严格的上凸函数。

如果 $f[\alpha\boldsymbol{x}_1+(1-\alpha)\boldsymbol{x}_2]\leqslant\alpha f(\boldsymbol{x}_1)+(1-\alpha)f(\boldsymbol{x}_2)$，则称 $f(\boldsymbol{x})$ 为定义域上的下凸函数，若式中“<”成立，则称为严格的下凸函数。

扩展阅读：信息增益与决策树

信息增益（Information Gain）通常用在决策树中，用来表示采用某种属性分类后不确定性的减少量，依据这个减少量衡量是否采用该属性进行分类：

$$\mathrm{IG}=H(X)-H(X\mid Y) \tag{1.43}$$

可以看到，它与平均互信息的计算公式是相同的，但是在平均互信息中，Y 表示一个随机变量，而在信息增益中 Y 表示某分类属性。

构造好的决策树的关键在于如何选择属性。最常用的分类属性选择指标就是信息增益。从候选属性中选择属性信息增益大的属性，其划分子集的纯度高（平均熵值小），分类能力强，因此选择信息增益最大的属性作为分类属性

例如，汤姆经常去看网球比赛，但网球比赛是否举行要根据天气情况而定。汤姆可不想白跑一趟。于是他准备根据自己记录的天气数据，设计一个决策树来判断去不去打球，如表 1.3 所示。目标属性 play 有 yes 和 no 两个值，根据其他属性来预测这个目标属性值。

表 1.3　关于目标属性 play（打球/不打球）的训练数据

outlook	temperature	humidity	windy	play
sunny	hot	high	FALSE	no
sunny	hot	high	TURE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TURE	no
overcast	cool	normal	TURE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TURE	yes
overcast	mild	high	TURE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TURE	no

式（1.43）中的 X 表示打球和不打球两种情况。只看最后一列，得到打球的概率是$\frac{9}{14}$，不打球的概率是$\frac{5}{14}$。因此在没有任何先验信息的情况下，系统的熵（不确定性）为

$$H(X) = -\frac{9}{14}\log\frac{9}{14} - \frac{5}{14}\log\frac{5}{14} = 0.94 \tag{1.44}$$

再统计每种候选属性下是否打球的频度，如表 1.4 所示。

表 1.4　每种候选属性下是否打球的统计数据

outlook			temperature			humidity			windy			play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2	3	hot	2	2	high	3	4	FALSE	6	2	9	5
overcast	4	0	mild	4	2	normal	6	1	TURE	3	3		
rainy	3	2	cool	3	1								

如果选 outlook 作为决策树的根节点，那么式（1.43）中的 Y 为集合{sunny, overcast, rainy}

Entropy(Sunny) = 0.971 比特/符号

$$\text{Entropy}(\text{Overcast}) = 0 \text{ 比特/符号}$$

$$\text{Entropy}(\text{Rain}) = 0.971 \text{ 比特/符号}$$

$$\text{IG} = H(X) - H(X \mid Y)$$

$$\text{Gain}(\text{Outlook}) = 0.940 - \left(\frac{5}{14}\right) \times \text{Entropy}(\text{Sunny}) - \left(\frac{4}{14}\right) \times \text{Entropy}(\text{Overcast}) - \left(\frac{5}{14}\right) \times \text{Entropy}(\text{Rain})$$

$$= 0.247 \text{ 比特/符号}$$

这里，第一个$\frac{5}{14}$是属性 Outlook 取值为 sunny 的个数占总记录的比例，同样$\frac{4}{14}$和第二个$\frac{5}{14}$是其取值为 overcast 和 rainy 的记录个数与总记录数之比。

因此，选择 outlook 作为决策树的根节点时，信息增益为 0.94 − 0.693 = 0.247 比特/符号。

同理，可计算出当选择 temperature、humidity、windy 作为根节点时系统的信息增益。

这样就得到了以上 4 个属性相应的信息增益值：

$$\text{Gain}(\text{Wind}) = 0.048 \text{ 比特/符号}$$

$$\text{Gain}(\text{Humidity}) = 0.151 \text{ 比特/符号}$$

$$\text{Gain}(\text{Outlook}) = 0.247 \text{ 比特/符号}$$

$$\text{Gain}(\text{Temperature}) = 0.029 \text{ 比特/符号}$$

选择 IG 值最大的 outlook 作为最终的根节点。

如果分类结果还不是单一类别，则再找另一属性继续划分，直至分类节点上的数据均对应于同一类别，并且所有属性都已使用过。

动手实践：图像的熵和平均互信息

编程实现计算两幅图像的信息熵以及它们的平均互信息。

输入：读入两幅图像。

输出：两幅图像的信息熵及它们之间的平均互信息。提示：概率分布根据灰度直方图求。

习　题　1

1.1　同时掷两枚骰子，事件 A、B、C 分别表示如下：A 表示仅有一枚骰子是 3，B 表示至少有一枚骰子是 4，C 表示骰子上点数的总和为偶数。试计算事件 A、B、C 发生后所提供的信息量。

1.2　设有 n 个球，每个球都以同样的概率$\frac{1}{N}$落入 N 个格子（$N \geqslant n$）的每个格子中。假定：A 表示某指定的 n 个格子中各落入一个球，B 表示任何 n 个格子中各落入一个球。试计算事件 A、B 发生后所提供的信息量。

1.3 一信源有4种输出符号 $x_i(i=0,1,2,3)$，且 $p(x_i)=\frac{1}{4}$。设信源向信宿发出 x_3，但由于传输中的干扰，接收者收到 x_3 后，认为其可信度为0.9。于是信源再次向信宿发送该符号（x_3），信宿无误收到。问信源在两次发送中发出的信息量各是多少？信宿在两次接收中得到的信息量又各是多少？

1.4 用递推性计算熵函数 $H\left(\frac{1}{3},\frac{1}{3},\frac{1}{6},\frac{1}{6}\right)$的值。

1.5 一个信源有6种输出状态，概率分别为

$p(A)=0.5$，$p(B)=0.25$，$p(C)=0.125$，$p(D)=p(E)=0.05$，$p(F)=0.025$

试计算 $H(X)$。然后求消息 *ABABBA* 和 *FDDFDF* 的信息量（设信源先后发出的符号相互独立），并将之与长度为6的消息序列信息量的期望值相比较。

1.6 中国国家标准局所规定的二级汉字共6763个。设每字使用的频度相等，求一个汉字所含的信息量。设每个汉字用一个 16×16 的二元点阵显示。试计算：显示方阵所能表示的最大信息，以及显示方阵的利用率。

1.7 已知信源发出 a_1 和 a_2 两种消息，且 $p(a_1)=p(a_2)=\frac{1}{2}$。此消息在二进制对称信道上传输，信道传输特性为 $p(b_1\mid a_1)=p(b_2\mid a_2)=1-\varepsilon,p(b_1\mid a_2)=p(b_2\mid a_1)=\varepsilon$。求互信息量 $I(a_1;b_1)$ 和 $I(a_1;b_2)$。

1.8 已知二维随机变量 XY 的联合概率分布 $p(x_iy_j):p(0,0)=p(1,1)=\frac{1}{8},p(0,1)=p(1,0)=\frac{3}{8}$，求 $H(X\mid Y)$。

1.9 X 和 Y 是 $\{0,1,2,3\}$ 上的独立、均匀分布的随机变量，求：

（1）$H(X+Y)$，$H(X-Y)$，$H(X\cdot Y)$

（2）$H(X+Y,X-Y)$，$H(X+Y,X\cdot Y)$

1.10 棒球比赛中大卫和麦克在前面的比赛中打平，最后三场与其他选手的比赛结果将最终决定他们的胜、负或平。

（1）假定最后三场他们与其他选手的比赛结果胜负的可能性均为0.5，把麦克的最终比赛结果{胜,负,平}作为随机变量，计算它的熵。

（2）计算麦克的最终比赛结果在假定大卫最后三场比赛全部获胜情况下的条件熵。

1.11 X、Y、Z 为三个随机变量，证明以下不等式成立并指出等号成立的条件：

（1）$H(XY\mid Z)\geqslant H(X\mid Z)$

（2）$I(XY;Z)\geqslant I(X;Z)$

（3）$H(XYZ)-H(XY)\leqslant H(XZ)-H(X)$

（4）$I(X;Z\mid Y)\geqslant I(Z;Y\mid X)-I(Z;Y)+I(X;Z)$

1.12 找出一个概率分布 $\{p_1,p_2,\cdots,p_5\}$，并且 $p_i>0$，使得 $H(p_1,p_2,\cdots,p_5)=2$。

1.13 有两个二元随机变量 X 和 Y，它们的联合概率分布如题表1.13所示。同时定义另一随机变量 $Z=X\cdot Y$（一般乘积）。试计算：

（1）熵 $H(X),H(Y),H(Z),H(XZ),H(YZ)$ 和 $H(XYZ)$。

（2）条件熵 $H(X\mid Y),H(Y\mid X)$，$H(X\mid Z)$，$H(Z\mid X)$，$H(Y\mid Z)$，$H(Z\mid Y)$，

$H(X|YZ)$，$H(Y|XZ)$和$H(Z|XY)$。

（3）互信息$I(X;Y)$，$I(X;Z)$，$I(Y;Z)$，$I(X;Y|Z)$，$I(Y;Z|X)$和$I(X;Z|Y)$。

1.14 假定$X_1 \to X_2 \to X_3 \to \cdots \to X_n$形成一个马尔可夫链，那么

$$p(x_1x_2\cdots x_n) = p(x_1)p(x_2|x_1)\cdots p(x_n|x_{n-1})$$

请化简$I(X_1;X_2\cdots X_n)$。

1.15 给定X、Y的联合概率分布如题表1.15所示。求：

题表1.13

Y \ X	0	1
0	$\frac{1}{8}$	$\frac{3}{8}$
1	$\frac{3}{8}$	$\frac{1}{8}$

题表1.15

X \ Y	0	1
0	$\frac{1}{3}$	$\frac{1}{3}$
1	0	$\frac{1}{3}$

（1）$H(X)$，$H(Y)$　　（2）$H(X|Y)$，$H(Y|X)$

（3）$H(XY)$　　（4）$H(Y)-H(Y|X)$

（5）$I(X;Y)$

1.16 （1）假定X是一个离散随机变量，$g(X)$是X的函数，证明：$H(g(X)) \leqslant H(X)$。

（2）假定X是一个定义在$\{0,1,2,3,4\}$上的等概分布的离散随机变量，$g(X)=\cos\frac{\pi X}{2}$，$f(X)=x^2$，比较它们的熵的大小。

1.17 考虑两个发射机和一个接收机之间的平均联合互信息$I(X_1X_2;Y)$。

（1）证明：$I(X_1X_2;Y) \geqslant I(X_1;Y)$，即用两台发射机比用一台发射机的效果好。

（2）证明：如果X_1和X_2相互独立，那么$I(X_2;Y|X_1) \geqslant I(X_2;Y)$。

（3）证明：如果X_1和X_2相互独立，那么$I(X_1X_2;Y) \geqslant I(X_1;Y)+I(X_2;Y)$，即同时用两台发射机比单独用两台发射机的效果好。

1.18 在一个布袋中有三枚硬币，分别用H、T、F表示，H的两面都是正面，T的两面都是反面，而F是一个一正一反的均匀硬币。随机选择一枚硬币并投掷两次，用X表示所选择的硬币，Y_1和Y_2表示两次投掷的结果，Z表示两次投掷中出现正面的次数。求：

（1）$I(X;Y_1)$　　（2）$I(X;Z)$　　（3）$I(Y_1;Y_2)$

1.19 猜宝游戏。三扇门中有一扇门后藏有一袋金子，并且三扇门后面藏有金子的可能性相同。如果有人随机打开一扇门并告诉你门后是否藏有金子，他给了你多少关于金子位置的信息？

1.20 一个年轻人研究了当地的天气纪录和气象台的预报纪录后，得到实际天气和预报天气的联合概率分布如题表1.20所示。他发现预报只有$\frac{12}{16}$的准确率，而不管三七二十一都预报明天不下雨的准确

题表1.20

预报 \ 实际	下雨	不下雨
下雨	$\frac{1}{8}$	$\frac{3}{16}$
不下雨	$\frac{1}{16}$	$\frac{10}{16}$

率却是$\frac{13}{16}$。他觉得没有必要预报了。他把这个想法跟气象台台长说了后，台长却说他错了。请问这是为什么？

1.21 设 $X_1, X_2, \cdots, X_N$ 为一个独立的贝努利随机变量序列，其分布为 $P_r\{X_i=0\}=p$，$P_r\{X_i=1\}=1-p$。求：

（1）使 $S_2=X_1+X_2$ 的熵 $H(S_2)$ 取得最大值的 p 值。

（2）设 $p=\frac{1}{2}$，求二项式随机变量 $S_n=X_1+X_2+\cdots+X_n$ 的熵 $H(S_n)$。

1.22 掷一枚均匀硬币直到出现两次正面或两次反面。用 X_1、X_2 分别表示头两次投掷，Y 表示最后一次投掷，N 表示投掷的次数。计算 $H(X_1)$，$H(X_2)$，$H(Y)$，$H(N)$，$I(X_1;Y)$，$I(X_2;Y)$，$I(N;Y)$，$I(X_1;N)$ 和 $I(X_2;N)$。

1.23 判断题。

（1）$H(X)>0$。

（2）若 X 与 Y 独立，则 $H(X)=H(X\mid Y)$。

（3）$I(X;Y)\geqslant I(X;Y\mid Z)$。

（4）若 $H(X\mid YZ)=0$，则要么 $H(X\mid Y)=0$，要么 $H(X\mid Z)=0$。

（5）$I(X;Y)\leqslant H(Y)$。

（6）$H(X\mid X)=0$。

（7）若 X 与 Y 独立，则 $H(Y\mid X)=H(X\mid Y)$。

（8）$H(X\mid Y)\geqslant H(X\mid YZ)$。

1.24 设随机变量 X 的概率分布为 $\left\{\frac{2}{10},\frac{2}{10},\frac{2}{10},\frac{1}{10},\frac{1}{10},\frac{1}{10},\frac{1}{10}\right\}$。随机变量 Y 是 X 的函数，其定义为将 X 的 4 个最小的概率分布合并为一个：$\left\{\frac{2}{10},\frac{2}{10},\frac{2}{10},\frac{4}{10}\right\}$。

（1）显然 $H(X)\leqslant\log_2 7$，请解释原因。

（2）请解释为什么 $H(X)>\log_2 5$。

（3）计算 $H(X)$ 和 $H(Y)$。

（4）计算 $H(X\mid Y)$ 并解释其结果。

1.25 已知 $H(Y\mid X)=0$，求证 $\forall x, p(x)>0$，只存在一个 y 使得 $p(xy)>0$。

1.26 猜宝游戏。三扇门中有一扇门后藏有一袋金子，并且三扇门后面藏有金子的可能性相同。你选择其中一扇门，主持人会打开后面没藏有金子的另一扇门（如果你选择的门后藏有金子，则主持人会在另两扇门中任意打开一扇门）。主持人给了你多少关于金子位置的信息？

1.27 在一个布袋中有 r 个红球、w 个白球、b 个黑球，从布袋中取 $k\geqslant 2$ 个球，问每次取出球后放回还是不放回的熵 $H(X_i\mid X_{i-1}\cdots X_1)$ 更大？

1.28 X、Y_1、Y_2 为二元随机变量，若 $I(X;Y_1)=0$ 且 $I(X;Y_2)=0$，能不能推出 $I(X;Y_1Y_2)=0$？如果能，请证明；如果不能，请给出反例。

1.29 X 是一个几何分布的随机变量，求它的熵。

1.30 人口问题。在某个地区，一对夫妻只允许生一个孩子，可是这里所有的夫妻都希望能生一个男孩传宗接代，因此这里的夫妻都会一直生到生了一个男孩为止。假定生

男生女的概率相同，问：

（1）这个地区男孩会多于女孩吗？

（2）一个家庭孩子的个数用离散随机变量 X 表示，计算 X 的熵。

1.31 就业问题。假如政府的就业问题顾问在考虑全国的就业问题时，把全体国民的就业情况分为三类：全就业（100% 就业）、部分就业（50% 就业）、失业（0% 就业），分别用概率 $p(E)$、$p(F)$、$p(U)$ 表示，要使全民的就业率达到 95%。请问：

（1）$p(E)$ 的取值范围。

（2）就业情况的熵作为 $p(E)$ 的函数，画出它在 $p(E)$ 的取值范围内的曲线。

（3）求就业情况熵的最大值。

第 2 章　信源及信源熵

信源（Information Source）是信息的来源，是产生消息（符号）、时间离散的消息序列（符号序列）以及时间连续的消息的来源。

信源输出的消息都是随机的，因此可用概率来描述其统计特性。在信息论中，用随机变量 X、随机矢量 $\boldsymbol{X}$、随机过程 $\{X(e,t)\}$ 分别表示产生消息、消息序列和时间连续消息的信源。

信源的主要问题包括：① 如何描述信源（信源的数学建模问题）；② 怎样定量描述信源输出信息的能力；③ 怎样有效地表示信源输出的消息，也就是信源编码问题。本章介绍前两个问题，重点是第二个问题，即计算信源输出信息的能力——熵率，在第 4、6 章将介绍第三个问题。下面分类介绍信源的数学模型及其熵率的计算。

2.1　信源的分类及其数学模型

第 1 章中已经介绍了离散随机变量及信息熵，离散随机变量表示信源输出的是一个符号的消息，如掷一颗骰子的试验，而通常实际信源输出的消息往往是时间（或空间）的函数，如掷多颗骰子的试验，消息的取值还可能是连续的，如多人跳远比赛的结果。

信源的分类有多种方法，我们常根据信源输出的消息在时间和取值上是离散的或连续的进行分类，如表 2.1 所示。

表 2.1　信源的分类

时间（空间）	取值	信源种类	举　例	数学描述
离散	离散	离散信源 （数字信源）	文字、数据、离散化图像	离散随机变量序列 $P(\boldsymbol{X})=P(X_1X_2\cdots X_N)$
离散	连续	连续信源	跳远比赛的结果、语音信号抽样以后	连续随机变量序列 $P(\boldsymbol{X})=P(X_1X_2\cdots X_N)$
连续	连续	波形信源 （模拟信源）	语音、音乐、热噪声、图形、图像	随机过程 $\{X(e,t)\}$
连续	离散		不常见	

实际信源输出的消息，如平时说话的语声和图像，在时间（或空间）和取值上都是连续的。这样的信源称为波形信源，用随机过程 $\{X(e,t)\}$ 描述。对于频率或时间受限的随机过程，根据抽样定理，人们通常把它转化成时间（或频率）离散的随机序列来处理，这样的信源称为连续信源。取样后的值通常还是连续的，因此还可以进一步经过分层量化，将连续随机变量转化成离散随机变量，连续信源变成离散信源来处理。

此外，根据各维随机变量的概率分布是否随时间的推移而变化，信源可以分为平稳信源和非平稳信源；根据随机变量间是否统计独立，信源可以分为有记忆信源和无记忆信源。

一个实际信源的统计特性往往是相当复杂的，要想找到精确的数学模型很困难。实际应

用时常常用一些可以处理的数学模型来近似。比如，语音信号就是非平稳随机过程，但人们常常用平稳随机过程来近似。平稳随机过程抽样以后的结果就是平稳随机序列。在数学上，随机序列是随机过程的一种，是时间参数离散的随机过程，这里把它单列出来。随机序列特别是离散平稳随机序列是我们研究的主要内容。实际信源的分类如下：

随机过程：波形信源
- 平稳信源
  - 离散平稳信源
    - 离散无记忆信源
    - 离散有记忆信源
      - 记忆长度无限长
      - 记忆长度有限（马尔可夫信源）
  - 连续平稳信源
- 非平稳信源

2.2 离散单符号信源

输出离散取值的单个符号的信源称为离散单符号信源，它是最简单也最基本的信源，是组成实际信源的基本单元，用一个离散随机变量表示。

信源所有可能输出的消息和消息对应的概率共同组成的二元序对$[X,P(X)]$被称为信源的概率空间，即

$$\begin{bmatrix} X \\ P(X) \end{bmatrix}=\begin{bmatrix} X=x_1 & \cdots & X=x_i & \cdots & X=x_q \\ p(x_1) & \cdots & p(x_i) & \cdots & p(x_q) \end{bmatrix}$$

其中，X 表示信源输出的消息的整体，x_i 表示某个消息，$p(x_i)$表示消息 x_i 出现的概率。q 是信源可能输出的消息数（信源可能输出的消息数可以是有限个，也可以是无限个，通常是有限个），这些消息两两互不相容，信源每次输出其中的一个消息。

注意，$p(x_i)$满足概率空间的非负性和完备性：

$$0\leqslant p(x_i)\leqslant 1$$

$$\sum_{i=1}^{q} p(x_i) = 1$$

信源输出的所有消息的自信息的统计平均值，定义为信源的平均自信息量（信息熵），它表示离散单符号信源的平均不确定性，即

$$H(X) = E[-\log p(x_i)] = -\sum_{i=1}^{q} p(x_i)\log p(x_i) \tag{2.1}$$

【例 2.1】二元信源$\begin{bmatrix} X \\ P(X) \end{bmatrix}=\begin{bmatrix} 0 & 1 \\ p & q \end{bmatrix}$，$p+q=1$，求 $H(X)$。

【解】$$\begin{aligned} H(X) &= -\sum_{i=1}^{q} p_i\log p_i \\ &= -p\log p-(1-p)\log(1-p) \\ &= H(p) \end{aligned}$$

$H(X)$是概率 p 的函数，通常用 $H(p)$表示，p 取值于$[0,1]$区间，如图 2.1 所示。

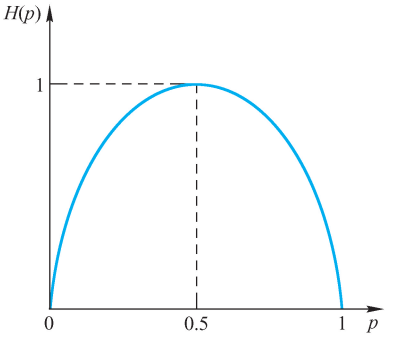


图 2.1 二元熵函数

若输出符号是确定的，即 $p=1$ 或 $p=0$，则 $H(p)=0$，信源不提供任何信息。若 $p=0.5$，即输出符号 0、1 以等概率发生时，信源熵达到极大值，平均每符号等于

1 比特信息量。

2.3 离散多符号信源

前面介绍的单符号信源是最简单的信源模型，用一个离散随机变量表示。实际信源输出的往往是符号序列，称为离散多符号信源，通常用离散随机变量序列（随机矢量）来表示：$\boldsymbol{X}=X_1X_2\cdots$。例如，电报系统发出的是一串有无脉冲的信号（用有脉冲表示 1，无脉冲表示 0），因此电报系统是输出一串 0、1 序列的二元信源。

为简单起见，这里我们只研究离散平稳信源，也就是统计特性不随时间改变的信源。下面先给出离散平稳信源的严格数学定义。

【定义 2-1】对于随机变量序列 $X_1,X_2,\cdots,X_n,\cdots$，在任意两个不同时刻 i 和 j（i 和 j 为大于 1 的任意整数），信源发出消息的概率分布完全相同，也就是对于任意 $N=0,1,2,\cdots$，$X_iX_{i+1}\cdots X_{i+N}\cdots$和 $X_jX_{j+1}\cdots X_{j+N}\cdots$具有相同的概率分布，即

$$P(X_i)=P(X_j) \tag{2.2}$$

$$P(X_iX_{i+1})=P(X_jX_{j+1}) \tag{2.3}$$

$$\vdots$$

$$P(X_iX_{i+1}\cdots X_{i+N})=P(X_jX_{j+1}\cdots X_{j+N}) \tag{2.4}$$

各维联合概率分布均与时间起点无关的信源称为离散平稳信源。

根据式(2.2)至式(2.4)以及联合概率与条件概率的关系，可得

$$P(X_{i+1}\mid X_i)=P(X_{j+1}\mid X_j) \tag{2.5}$$

$$\vdots$$

$$P(X_{i+N}\mid X_iX_{i+1}\cdots X_{i+N-1})=P(X_{j+N}\mid X_jX_{j+1}\cdots X_{j+N-1}) \tag{2.6}$$

即离散平稳信源的条件概率分布均与时间起点无关，而只与关联长度 N 有关。这样我们很容易推出

$$H(X_1)=H(X_2)=\cdots=H(X_N) \tag{2.7}$$

$$H(X_2\mid X_1)=H(X_3\mid X_2)=\cdots=H(X_N\mid X_{N-1}) \tag{2.8}$$

$$H(X_3\mid X_1X_2)=H(X_4\mid X_2X_3)=\cdots=H(X_N\mid X_{N-2}X_{N-1}) \tag{2.9}$$

$$\vdots$$

对于离散单符号信源，我们用信息熵来表示信源的平均不确定性。对于离散多符号信源，怎样表示信源的平均不确定性呢？下面引入“熵率”的概念，它表示信源输出的符号序列中，平均每个符号所携带的信息量。

【定义 2-2】随机变量序列中，对前 N 个随机变量的联合熵求平均称为平均符号熵，即

$$H_N(\boldsymbol{X})=\frac{1}{N}H(X_1X_2\cdots X_N) \tag{2.10}$$

如果 $N\to\infty$ 时上式的极限存在，则 $\lim\limits_{N\to\infty}H_N(\boldsymbol{X})$ 称为熵率，或称为极限熵，记为

$$H_\infty\overset{\text{def}}{=}\lim_{N\to\infty}H_N(\boldsymbol{X}) \tag{2.11}$$

2.3.1 离散平稳无记忆信源

一般情况下，信源输出序列中的每一位出现什么符号是随机的，但是前后符号的出现有

一定的统计关系。为简单起见，我们先假定消息符号序列中前后符号的出现是无关的，即我们首先讨论无记忆信源。

离散平稳无记忆信源输出的符号序列是平稳随机序列，并且符号之间是无关的，即统计独立的。为了研究离散平稳无记忆信源的熵率，我们假定信源每次输出的是 N 长的符号序列，这可视为一个新信源，称为离散平稳无记忆信源的 N 次扩展信源，它的数学模型是 N 维离散随机变量序列（随机矢量）$\boldsymbol{X}=X_1X_2\cdots X_N$，其中每个随机变量之间统计独立。同时，由于是平稳信源，每个随机变量的统计特性都相同，我们还可以把 N 次扩展信源的输出记为 $\boldsymbol{X}=X_1X_2\cdots X_N=X^N$。

根据统计独立的多维随机变量的联合熵与信息熵之间的关系，可以推出

$$H(\boldsymbol{X})=H(X^N)=NH(X) \tag{2.12}$$

即 N 次扩展信源的熵等于单符号离散信源熵的 N 倍，信源输出的 N 长符号序列平均提供的信息量，是单符号离散信源平均每个符号所提供信息量的 N 倍。这似乎很好理解，如抛掷一枚均匀硬币的试验每次可以得到 1 比特的信息量，抛掷 N 枚均匀硬币的试验则可以得到 N 比特的信息量。

离散平稳无记忆信源的熵率

$$H_\infty=\lim_{N\to\infty}H_N(\boldsymbol{X})=\lim_{N\to\infty}\frac{1}{N}\cdot NH(X)=H(X) \tag{2.13}$$

【例 2.2】设有一离散无记忆信源 X，其概率空间为

$$\begin{bmatrix} X \\ P(X) \end{bmatrix}=\begin{bmatrix} x_1 & x_2 & x_3 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$$

求该信源的熵率及其二次扩展信源（信源每次输出两个符号）的熵。

【解】

单符号离散信源熵为

$$\begin{aligned} H(X) &= -\sum_{i=1}^{q}p_i\log p_i \\ &=\frac{1}{2}\log 2+\frac{1}{4}\log 4+\frac{1}{4}\log 4 \\ &=1.5\ \text{比特/符号} \end{aligned}$$

熵率为

$$\begin{aligned} H_\infty &= \lim_{N\to\infty}H_N(\boldsymbol{X}) \\ &=\lim_{N\to\infty}\frac{1}{N}\cdot NH(X) \\ &=H(X) \\ &=1.5\ \text{比特/符号} \end{aligned}$$

二次扩展信源的熵为

$$H(\boldsymbol{X})=2H(X)=3\ \text{比特/二个符号}$$

注意，$H(\boldsymbol{X})$ 的单位是“比特/二个符号”，其中每个符号提供的信息量仍然是 1.5 比特。

2.3.2 离散平稳有记忆信源

前面介绍了离散平稳信源中最简单的离散平稳无记忆信源，而实际信源往往是有记忆信源。假定信源输出 N 长的符号序列，则它的数学模型是 N 维离散随机变量序列（随机矢量）$\boldsymbol{X}=X_1X_2\cdots X_N$，其中每个随机变量之间存在统计依赖关系。

相互间有依赖关系的 N 维随机变量的联合熵可以用式(2.14)表示，这称为熵函数的链规则：

$$\begin{aligned}H(\boldsymbol{X})&=H(X_1X_2\cdots X_N)\\&=H(X_1)+H(X_2\mid X_1)+H(X_3\mid X_1X_2)+\cdots+H(X_N\mid X_1X_2\cdots X_{N-1})\end{aligned}\tag{2.14}$$

N 维随机变量的联合熵等于起始时刻随机变量 X_1 的熵与各阶条件熵之和。

【定理 2-1】对于离散平稳信源，有以下几个结论：

（1）条件熵 $H(X_N\mid X_1X_2\cdots X_{N-1})$ 随 N 的增加是递减的。

（2）N 给定时，平均符号熵≥条件熵，即

$$H_N(\boldsymbol{X})\geqslant H(X_N\mid X_1X_2\cdots X_{N-1})\tag{2.15}$$

（3）平均符号熵 $H_N(\boldsymbol{X})$ 随 N 的增加是递减的。

（4）如果 $H(X_1)<\infty$，则 $H_\infty=\lim\limits_{N\to\infty}H_N(\boldsymbol{X})$ 存在，并且

$$H_\infty=\lim_{N\to\infty}H_N(\boldsymbol{X})=\lim_{N\to\infty}H(X_N\mid X_1X_2\cdots X_{N-1})\tag{2.16}$$

【证明】

（1）$\begin{aligned}H(X_N\mid X_1X_2\cdots X_{N-1})&\leqslant H(X_N\mid X_2\cdots X_{N-1})\quad\text{（条件熵小于等于无条件熵）}\\&=H(X_{N-1}\mid X_1X_2\cdots X_{N-2})\quad\text{（序列的平稳性）}\end{aligned}$

所以，条件熵 $H(X_N\mid X_1X_2\cdots X_{N-1})$ 随着 N 的增加是递减的。这表明记忆长度越长，条件熵越小，也就是序列的统计约束关系增加时，不确定性减少。

（2）$\begin{aligned}NH_N(\boldsymbol{X})&=H(X_1X_2\cdots X_N)\\&=H(X_1)+H(X_2\mid X_1)+H(X_3\mid X_1X_2)+\cdots+H(X_N\mid X_1X_2\cdots X_{N-1})\\&=H(X_N)+H(X_N\mid X_{N-1})+\cdots+H(X_N\mid X_1X_2\cdots X_{N-1})\quad\text{（序列的平稳性）}\\&\geqslant NH(X_N\mid X_1X_2\cdots X_{N-1})\quad\text{（条件熵小于无条件熵）}\end{aligned}$

所以，$H_N(\boldsymbol{X})\geqslant H(X_N\mid X_1X_2\cdots X_{N-1})$。即 N 给定时，平均符号熵≥条件熵。

（3）$\begin{aligned}NH_N(\boldsymbol{X})&=H(X_1X_2\cdots X_N)=H(X_N\mid X_1X_2\cdots X_{N-1})+H(X_1X_2\cdots X_{N-1})\\&=H(X_N\mid X_1X_2\cdots X_{N-1})+(N-1)H_{N-1}(\boldsymbol{X})\\&\leqslant H_N(\boldsymbol{X})+(N-1)H_{N-1}(\boldsymbol{X})\quad[\text{利用式(2.15)}]\end{aligned}$

所以，$H_N(\boldsymbol{X})\leqslant H_{N-1}(\boldsymbol{X})$。即序列的统计约束关系增加时，由于符号间的相关性，平均每个符号所携带的信息量减少。

（4）只要 X_1 的样本空间是有限的，则必然有 $H(X_1)<\infty$，因此

$$0\leqslant H(X_N\mid X_1X_2\cdots X_{N-1})\leqslant H(X_{N-1}\mid X_1X_2\cdots X_{N-2})\leqslant\cdots\leqslant H(X_1)<\infty$$

从而 $H(X_N\mid X_1X_2\cdots X_{N-1})(N=1,2,\cdots)$ 是一个单调有界数列，极限 $\lim\limits_{N\to\infty}H(X_N\mid X_1X_2\cdots X_{N-1})$

必然存在，且极限为 0 和 $H(X_1)$ 之间的某一值。

对于收敛的实数列有以下等式成立：如果 $a_1, a_2, a_3, \cdots$ 是一个收敛的实数列，那么

$$\lim_{N\to\infty}\frac{1}{N}(a_1+a_2+\cdots+a_N)=\lim_{N\to\infty}a_N \tag{2.17}$$

利用式(2.17)可以推出

$$\begin{aligned}\lim_{N\to\infty}H_N(\boldsymbol{X}) &= \lim_{N\to\infty}\frac{1}{N}[H(X_1)+H(X_2\mid X_1)+H(X_3\mid X_1X_2)+\cdots+H(X_N\mid X_1X_2\cdots X_{N-1})]\\ &= \lim_{N\to\infty}H(X_N\mid X_1X_2\cdots X_{N-1})\end{aligned}$$

证毕。

定理 2-1 表明，由于信源输出序列前后符号之间的统计依存关系，随着 N 的增加，即统计约束条件不断增加，平均符号熵 $H_N(\boldsymbol{X})$ 及条件熵 $H(X_N\mid X_1X_2\cdots X_{N-1})$ 均随之减小。当 $N\to\infty$ 时，$H_N(\boldsymbol{X})=H(X_N\mid X_1X_2\cdots X_{N-1})$，即为熵率，它表示信源输出的符号序列中，平均每个符号所携带的信息量。所以，求离散平稳有记忆信源的熵率时可以有两种途径，可以求它的极限平均符号熵，也可以求它的极限条件熵：

$$\begin{aligned}H_\infty &= \lim_{N\to\infty}\frac{1}{N}H(X_1X_2\cdots X_N)\\ &= \lim_{N\to\infty}H(X_N\mid X_1X_2\cdots X_{N-1})\end{aligned}$$

一般情况下，平稳信源输出的符号序列中，符号之间的相关性可以追溯到最初的一个符号，如一篇文章的最后一句话可以一直追溯到与开篇第一句话相关。要准确地计算出这个熵率，必须测定信源的无穷维联合概率和条件概率分布，这相当困难。为简化分析，往往用 N 不太大时的平均符号熵或条件熵作为熵率的近似值。比如，英文信源的熵率通常用 $N=5$ 时的条件熵近似。

有一类信源在某时刻发出的符号仅与在此之前发出的有限个符号有关，而与更早些时候发出的符号无关，这类信源称为马尔可夫信源。马尔可夫信源可以在 N 不是很大时得到 H_∞。如果信源在某时刻发出的符号仅与在此之前发出的 m 个符号有关，则称为 m 阶马尔可夫信源，它的熵率为

$$\begin{aligned}H_\infty &= \lim_{N\to\infty}H(X_N\mid X_1X_2\cdots X_{N-1})\\ &= \lim_{N\to\infty}H(X_N\mid X_{N-m}X_{N-m+1}\cdots X_{N-1}) \quad (\text{马尔可夫性})\\ &= H(X_{m+1}\mid X_1X_2\cdots X_m) \quad (\text{序列的平稳性})\end{aligned} \tag{2.18}$$

式中，$H(X_{m+1}\mid X_1X_2\cdots X_m)$ 通常记为 H_{m+1}。

【例 2.3】信源 X 的信源模型为

$$\begin{bmatrix}X\\P(X)\end{bmatrix}=\begin{bmatrix}x_1 & x_2 & x_3\\ \frac{1}{4} & \frac{4}{9} & \frac{11}{36}\end{bmatrix}$$

输出符号序列中只有前后两个符号有记忆。条件概率 $P(X_2\mid X_1)$ 列于表 2.2 中。求熵率，并比较 $H(X_2\mid X_1)$、$\frac{1}{2}H(X_1X_2)$ 和 $H(X)$ 的大小。

表 2.2　条件概率 $P(X_2 \mid X_1)$

X_1 \ X_2	x_1	x_2	x_3
x_1	$\frac{7}{9}$	$\frac{2}{9}$	0
x_2	$\frac{1}{8}$	$\frac{3}{4}$	$\frac{1}{8}$
x_3	0	$\frac{2}{11}$	$\frac{9}{11}$

【解】

熵率为

$$H_\infty = H_{m+1} = H(X_2 \mid X_1)$$
$$= 0.870 \text{ 比特/符号}$$

如果不考虑符号间的相关性，则信源熵为

$$H(X) = \frac{1}{4}\log 4 + \frac{4}{9}\log\frac{9}{4} + \frac{11}{36}\log\frac{36}{11}$$
$$= 1.542 \text{ 比特/符号}$$

可见，$H(X_2 \mid X_1) < H(X) = H(X_2)$。这是由于 X_1 和 X_2 之间存在统计依赖关系，在 X_1 已知的情况下，X_2 的不确定度减少，即条件熵 $H(X_2 \mid X_1)$ 小于无条件熵 $H(X_2)$。因此，在考虑序列符号之间的相关性之后，序列的熵减小。

如果把信源输出的符号序列看成是分组发出的，每两个符号作为一组，这样可以把符号序列视为由一个新信源发出的，新信源每次发出的是由两个符号构成的消息。新信源的数学模型是一个二维随机变量，新信源的熵为

$$H(X_1X_2) = H(X_1) + H(X_2 \mid X_1)$$
$$= 1.542 + 0.870$$
$$= 2.412 \text{ 比特/二个符号}$$

平均符号熵为

$$\frac{1}{2}H(X_1X_2) = 1.206 \text{ 比特/符号}$$

可见

$$H(X_2 \mid X_1) < \frac{1}{2}H(X_1X_2) < H(X)$$

这是因为 $H(X_1X_2)$ 考虑了同一组的两个符号之间的相关性，所以 $H(X_1X_2)$ 小于不考虑符号间的相关性时的信源熵 $H(X)$，但 $H(X_1X_2)$ 未考虑前一组的后一符号与后一组的前一符号之间的关联，所以

$$H(X_2 \mid X_1) < \frac{1}{2}H(X_1X_2)$$

2.3.3　马尔可夫信源

前面讨论了离散平稳信源的熵率，由于符号间的相关性可以追溯到很远，使得熵率的计算比较复杂。马尔可夫信源是一类相对简单的有记忆信源，信源在某一时刻发出某一符号的

概率除与该符号有关外，只与此前发出的有限个符号有关。例如，m 阶马尔可夫信源只与前面发出的 m 个符号有关，而 1 阶马尔可夫信源只与前面 1 个符号有关。因此，如果把前面若干个符号视为一个状态（若信源有 q 个可能的输出符号，则一共有 q^m 个可能的状态），那么可以认为，信源在某一时刻发出某一符号的概率除了与该符号有关外，只与该时刻信源所处的状态有关，而与过去的状态无关。信源发出一个符号后，信源所处的状态即发生改变，这些状态的变化组成了马尔可夫链。因此我们可以把对马尔可夫信源的研究转化对马尔可夫链的研究。

如图 2.2 所示，信源在某时刻处于某状态 s_i，当它发出一个符号 $x_{i_{m+1}}$ 后，所处状态就变了，转移到状态 s_j，因此信源输出的符号序列 $X_1X_2\cdots X_mX_{m+1}\cdots$ 变换成信源状态序列 $S_1S_2\cdots S_LS_{L+1}\cdots$，于是讨论信源输出符号的不确定性问题变成了讨论信源状态转移的问题。

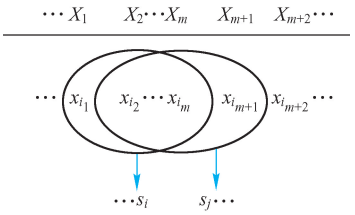


图 2.2　马尔可夫信源

状态之间的一步转移概率 $p_{ij}=P(S_{L+1}=s_j\mid S_L=s_i)$ 表示前一时刻即 L 时刻信源处于 s_i 状态下，在下一时刻即 $L+1$ 时刻信源处于 s_j 状态的概率。用马尔可夫链的状态转移图可以方便地描述离散马尔可夫信源的状态转移概率。

【例 2.4】设有一个二元一阶马尔可夫信源，信源符号集为 $X=\{0,1\}$，输出符号的条件概率为 $p(0\mid 0)=0.25$，$p(0\mid 1)=0.5$，$p(1\mid 0)=0.75$，$p(1\mid 1)=0.5$，求状态转移概率。

【解】

这里 $q=2$，$m=1$，$q^m=2$，信源有两种状态：$s_1=0$，$s_2=1$。

由输出符号的条件概率，可求得信源的状态转移概率：

$$p(s_1\mid s_1)=0.25$$
$$p(s_1\mid s_2)=0.5$$
$$p(s_2\mid s_1)=0.75$$
$$p(s_2\mid s_2)=0.5$$

信源的状态转移概率还可以用如图 2.3 所示的状态转移图表示。

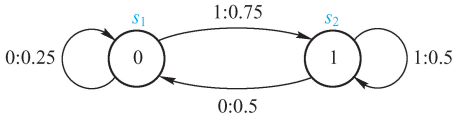


图 2.3　一阶马尔可夫信源状态转移图

对于一阶马尔可夫信源，它的状态转移概率和信源输出符号的条件概率即符号转移概率相同。

【例 2.5】设有一个二元二阶马尔可夫信源，其信源符号集为 $X=\{0,1\}$，输出符号的条件概率为 $p(0\mid 00)=p(1\mid 11)=0.8$，$p(0\mid 01)=p(0\mid 10)=p(1\mid 01)=p(1\mid 10)=$

0.5，$p(1\mid 00)=p(0\mid 11)=0.2$，求状态转移概率矩阵。

【解】

这里 $q=2$，$m=2$，故信源共有 $q^m=4$ 个可能的状态：$s_1=00$，$s_2=01$，$s_3=10$，$s_4=11$。由于信源每次只可能发出 0 或 1，所以信源下一时刻只可能转移到其中的两种状态之一。比如，如果信源原来所处状态为 $s_1=00$，则下一时刻信源只可能转移到 00 或 01 状态，而不会转移到 10 或 11 状态。

由输出符号的条件概率，容易求得状态转移概率：

$$p(s_1\mid s_1)=p(s_4\mid s_4)=0.8$$

$$p(s_2\mid s_1)=p(s_3\mid s_4)=0.2$$

$$p(s_3\mid s_2)=p(s_1\mid s_3)=p(s_4\mid s_2)=p(s_2\mid s_3)=0.5$$

其余状态转移概率为 0，该信源的状态转移图如图 2.4 所示。信源的状态转移概率还可以用状态转移矩阵表示：

$$\boldsymbol{P}=\begin{bmatrix}0.8 & 0.2 & 0 & 0\\ 0 & 0 & 0.5 & 0.5\\ 0.5 & 0.5 & 0 & 0\\ 0 & 0 & 0.2 & 0.8\end{bmatrix}$$

解毕。

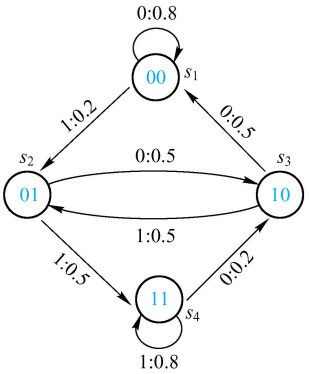


图 2.4　二阶马尔可夫信源状态转移图

对于一般的 m 阶马尔可夫信源，它的所有可能的输出符号及输出符号的条件概率可以组成马尔可夫信源的概率空间：

$$\begin{bmatrix}X\\ P(X)\end{bmatrix}=\begin{bmatrix}x_1 & \cdots & x_i & \cdots & x_q\\ p(x_{i_{m+1}}\mid x_{i_1}x_{i_2}\cdots x_{i_m})\end{bmatrix}$$

令 $s_i=x_{i_1}x_{i_2}\cdots x_{i_m}$（$i_1,i_2,\cdots,i_m\in\{1,2,\cdots,q\}$），则由信源输出符号的条件概率 $p(x_{i_{m+1}}\mid x_{i_1}x_{i_2}\cdots x_{i_m})$ 可以确定状态转移概率 $p(s_j\mid s_i)$（$i,j\in\{1,2,\cdots,q^m\}$），从而得到马尔可夫信源的状态空间：

$$\begin{bmatrix}s_1 & \cdots & s_i & \cdots & s_{q^m}\\ & & p(s_j\mid s_i) & & \end{bmatrix}$$

状态空间由所有状态及状态间的状态转移概率组成。因此通过引入状态转移概率，可以把对马尔可夫信源的研究转化为对马尔可夫链的研究。

我们主要研究遍历的 m 阶马尔可夫信源的熵率。

当时间足够长后，遍历的马尔可夫信源可视为平稳信源来处理，又因为 m 阶马尔可夫信源发出的符号只与最近的 m 个符号有关，所以

$$\begin{aligned}H_\infty&=\lim_{N\to\infty}H(X_N\mid X_1X_2\cdots X_{N-1})\\&=\lim_{N\to\infty}H(X_N\mid X_{N-m}X_{N-m+1}\cdots X_{N-1})\quad\text{（马尔可夫性）}\\&=H(X_{m+1}\mid X_1X_2\cdots X_m)\quad\text{（序列的平稳性）}\\&=H_{m+1}\end{aligned}\tag{2.19}$$

即 m 阶马尔可夫信源的极限熵 H_∞ 等于条件熵 H_{m+1}。H_{m+1} 表示已知前面 m 个符号的条件下，

输出下一个符号的平均不确定性。

对于齐次遍历的马尔可夫链，其状态 s_i 由 $x_{i_1}x_{i_2}\cdots x_{i_m}$ 唯一确定，因此

$$p(x_{i_{m+1}} \mid x_{i_1}x_{i_2}\cdots x_{i_m}) = p(x_{i_{m+1}} \mid s_i) = p(s_j \mid s_i) \tag{2.20}$$

所以

$$\begin{aligned} H_{m+1} &= H(X_{m+1} \mid X_1X_2\cdots X_m) \\ &= E[p(x_{i_{m+1}} \mid x_{i_1}x_{i_2}\cdots x_{i_m})] \\ &= E[p(x_{i_{m+1}} \mid s_i)] \\ &= -\sum_{i=1}^{q^m}\sum_{i_{m+1}=1}^{q} p(s_i)p(x_{i_{m+1}} \mid s_i)\log p(x_{i_{m+1}} \mid s_i) \\ &= \sum_i p(s_i)H(X \mid s_i) \qquad (2.21) \\ &= -\sum_i\sum_j p(s_i)p(s_j \mid s_i)\log p(s_j \mid s_i) \qquad (2.22) \end{aligned}$$

式中，$p(s_i)$是马尔可夫链的平稳分布或称状态极限概率，$H(X \mid s_i)$表示信源处于某一状态 s_i 时发出下一个符号的平均不确定性，$p(s_j \mid s_i)$是状态的一步转移概率。

【例 2.6】求图 2.4 中的二阶马尔可夫信源的极限熵。

【解】

根据马尔可夫链状态的分类方法可以判断，图 2.4 中的 4 个状态是不可约的非周期常返态，因此是遍历的。设状态的平稳分布为 $\boldsymbol{W} = [W_1 \quad W_2 \quad W_3 \quad W_4]$，其中 $W_1 = p(s_1)$，$W_2 = p(s_2)$，$W_3 = p(s_3)$，$W_4 = p(s_4)$，遍历的马尔可夫链满足方程 $\boldsymbol{WP} = \boldsymbol{W}$，即

$$\begin{cases} 0.8W_1 + 0.5W_3 = W_1 \\ 0.2W_1 + 0.5W_3 = W_2 \\ 0.5W_2 + 0.2W_4 = W_3 \\ 0.5W_2 + 0.8W_4 = W_4 \end{cases}$$

并且满足

$$W_1 + W_2 + W_3 + W_4 = 1$$

因此，可解得 $W_1 = p(s_1) = \frac{5}{14}$，$W_2 = p(s_2) = \frac{1}{7}$，$W_3 = p(s_3) = \frac{1}{7}$，$W_4 = p(s_4) = \frac{5}{14}$。

所以

$$\begin{aligned} H_\infty &= H_3 = \sum_i p(s_i)H(X \mid s_i) \\ &= \frac{5}{14}H(0.8,0.2) + \frac{1}{7}H(0.5,0.5) + \frac{1}{7}H(0.5,0.5) + \frac{5}{14}H(0.8,0.2) \\ &= 0.80 \text{ 比特/符号} \end{aligned}$$

注意，这时符号的平稳概率分布为

$$\begin{aligned} p(0) &= 0.8p(s_1) + 0.5p(s_2) + 0.5p(s_3) + 0.2p(s_4) = 0.5 \\ p(1) &= 0.2p(s_1) + 0.5p(s_2) + 0.5p(s_3) + 0.8p(s_4) = 0.5 \end{aligned}$$

它与状态的稳定分布是有区别的。

如果不考虑符号间的相关性，则由符号的平稳概率分布可得信源熵 $H(X) = 1$ 比特/符

号，而考虑符号间的相关性后，该信源的熵率为

$$H_\infty = H_{m+1} = H_3 = 0.80 \text{ 比特/符号}$$

2.3.4 信源的相关性和剩余度

前面讨论了离散平稳信源及其熵率。实际的离散信源可能是非平稳的，对于非平稳信源来说，其 H_∞ 不一定存在，但为了方便，通常假定它是平稳的，用平稳信源的 H_∞ 来近似。对于一般的离散平稳信源，求 H_∞ 值也是很困难的，那么进一步假定它是 m 阶马尔可夫信源，用 m 阶马尔可夫信源的条件熵 H_{m+1} 来近似（大多数平稳信源可用马尔可夫信源来近似，即认为输出符号只与前面的有限个符号有关）。$m=1$ 是最简单的离散平稳有记忆信源，这时 $H_{m+1} = H_2 = H(X_2 \mid X_1)$。若再进一步简化信源模型，则可以假设信源为离散平稳无记忆信源，这时可用单符号离散信源的平均自信息量来近似，$H_1 = H(X)$。最后，可以假定信源输出的符号是等概分布的，因此可以用最大离散熵来近似，$H_0 = \log q$。所以，对于一般的离散信源，根据我们研究的目的不同，可以用不同的信源模型来近似。

由定理 2-1 可知

$$\log q = H_0 \geqslant H_1 \geqslant H_2 \geqslant \cdots \geqslant H_{m+1} \geqslant \cdots \geqslant H_\infty$$

对于一个信源，其输出的每个符号实际所携带的平均信息量用熵率 H_∞ 表示。由于信源输出符号间的依赖关系也就是信源的相关性，使信源的 H_∞ 减小，信源输出符号间统计约束关系越长，信源的 H_∞ 越小。当信源输出符号间彼此不存在依赖关系且为等概分布时，信源的 H_∞ 等于最大熵 H_0。例如，信源符号集有 4 个符号，最大熵为 2 比特/符号，输出一个由 10 个符号构成的符号序列，最多包含 $10 \times 2 = 20$ 比特的信息量。假如，由于符号间的相关性或不等概分布，使信源的 H_∞ 减小到 1.2 比特/符号，则输出的符号序列平均所含有的信息量为 $10 \times 1.2 = 12$ 比特，如果信源输出符号间没有相关性并且符号等概分布，则输出 12 比特的信息量只需输出 6 个符号就可以了，说明信源存在剩余。因此，我们引入信源剩余度（冗余度）的概念。

【定义 2-3】 一个信源的熵率（极限熵）与具有相同符号集的最大熵的比值称为熵的相对率，即

$$\eta = \frac{H_\infty}{H_0} \tag{2.23}$$

信源剩余度为

$$\gamma = 1 - \eta = 1 - \frac{H_\infty}{H_0} = 1 - \frac{H_\infty}{\log q} \tag{2.24}$$

$H_0 - H_\infty$ 越大，信源的剩余度越大。

信源的剩余度来自两方面：一方面是信源符号间的相关性，相关程度越大，符号间的依赖关系越长，信源的 H_∞ 越小；另一方面是信源符号分布的不均匀性使信源的 H_∞ 减小。当信源输出符号间不存在相关性并且符号为等概分布时，信源的 H_∞ 最大，等于 H_0。一般，平稳信源的极限熵 H_∞ 远小于 H_0。传输一个信源的信息实际只需传输的信息量为 H_∞，如果用二元符号来表示，只需用 H_∞ 个二元符号。

为了最有效地传输信源的信息，就需要掌握信源全部的概率统计特性，即任意 N 维的

概率分布，这显然是不现实的。实际上，往往只能掌握有限 N 维的概率分布，这时传输的信息量为 H_N，因此与理论值 H_∞ 相比，就多传输了 $H_N - H_\infty$。

为了更经济有效地传输信息，需要尽量压缩信源的剩余度，压缩剩余度的方法是尽量减小符号间的相关性，并且尽可能地使信源符号等概分布。第 4 章无失真信源编码中将研究具体的信源剩余度压缩方法。

下面以英文字母为例来说明，信源模型的近似程度不同，计算的信源熵不同。

① 英文字母共 26 个，加上空格 27 个符号。最大熵 $H_0 = \log q = 4.76$ 比特/符号。

② 对在英文书中各字母（包括空格）出现的概率加以统计，不考虑字母之间的依赖关系，可以得到每个字母的概率分布，如表 2.3 所示。

表 2.3　英文字母概率表

字　母	P_i	字　母	P_i	字　母	P_i
空格	0.2	S	0.0502	Y/W	0.012
E	0.105	H	0.047	G	0.011
T	0.072	D	0.035	B	0.0105
O	0.0654	L	0.029	V	0.008
A	0.063	C	0.023	K	0.003
N	0.059	F/U	0.0225	X	0.002
I	0.055	M	0.021	J/Q	0.001
R	0.054	P	0.0175	Z	0.001

因此，如果认为英语字母间是无记忆的，则根据表中的概率可求得

$$H_1 = -\sum_{i=1}^{q} p(x_i)\log p(x_i)$$
$$= 4.03 \text{ 比特 / 符号}$$

③ 若考虑前后两个、三个、若干个字母之间存在相关性，则可根据字母出现的条件概率求得

$$H_2 = 3.32 \text{ 比特/符号}$$
$$H_3 = 3.1 \text{ 比特/符号}$$
$$\vdots$$
$$H_5 = 1.65 \text{ 比特/符号}$$
$$H_\infty = 1.4 \text{ 比特/符号（利用统计推断方法）}$$

如果考虑 5 个字母间的相关性（约等于英文单词的平均长度 4.5），所计算的信源熵已非常接近英文符号的实际信源熵 H_∞。

$$\eta = \frac{H_\infty}{H_0} = \frac{1.4}{4.76} = 0.29$$
$$\gamma = 1 - \eta = 0.71$$

这说明，写英语文章时，71% 是由语言结构定好的，是多余成分，只有 29% 是写文章的人可以自由选择的。可以这样理解，100 页英文书理论上看仅有 29 页是有效的，其余 71 页是多余的。正是由于这一多余量的存在，才有可能对英文信源进行压缩编码。如果对英文信源进行恰当的编码，传输或存储这些符号时可大量压缩篇幅，100 页的英语大约只要 29 页即可。

表 2.4 给出了 5 种文字在不同近似程度下的熵。

表 2.4　5 种文字在不同近似程度下的熵

文字	H_0	H_1	H_2	H_3	…	H_∞	η	γ
英文	4.7	4.03	3.32	3.1		1.4	0.29	0.71
法文	4.7					3	0.63	0.37
德文	4.7					1.08	0.23	0.77
西班牙文	4.7					1.97	0.42	0.58
中文（按 8000 汉字计算）	≈13	9.41	8.1	7.7		4.1	0.315	0.685

【例 2.7】计算汉字的剩余度。假设常用汉字约为 10000 个，其中 140 个汉字出现的概率占 50%，625 个汉字（含 140 个）出现的概率占 85%，2400 个汉字（含 625 个）出现的概率占 99.7%，其余 7600 个汉字出现的概率占 0.3%，不考虑符号间的相关性，只考虑它的概率分布，在这一级近似下计算汉字的剩余度。

【解】为了计算方便，假设每类中汉字出现是等概的，得到表 2.5。

表 2.5　汉字的近似概率表

类　别	汉字个数	所占概率	每个汉字的概率
1	140	0.5	$\frac{0.5}{140}$
2	625 − 140 = 485	0.85 − 0.5 = 0.35	$\frac{0.35}{485}$
3	2400 − 625 = 1775	0.997 − 0.85 = 0.147	$\frac{0.147}{1775}$
4	7600	0.003	$\frac{0.003}{7600}$

不考虑符号间的相关性，只考虑它的概率分布，因此信源的实际熵近似为 $H(X)=9.773$ 比特/汉字，$H_0=13.288$ 比特/汉字，从而 $\gamma=1-\frac{H(X)}{H_0}=0.264$。

从提高信息传输效率的观点出发，人们总是希望尽量去掉剩余度。比如发电报，人们都知道尽可能把电文写得简洁些以去除相关性，如“母病愈”三个字的中文电报就可以表达母亲身体情况好转的消息。但是从提高抗干扰能力角度来看，却希望增加或保留信源的剩余度，因为剩余度大的消息抗干扰能力强。比如，收到电文“母亲病 X，身体健康”，人们很容易把电文纠正为“母亲病愈，身体健康”，而收到电文“母病 X”我们就不知道对方发的是“母病愈”还是“母病危”。

本书从第 4 章开始将讨论信源编码和信道编码，我们可以进一步理解：信源编码减少或消除信源的剩余度，以提高信息的传输效率，信道编码则通过增加冗余度来提高信息传输的抗干扰能力。

2.4　连续信源

连续随机变量的取值是连续的，一般用概率密度函数来描述其统计特征。

单变量连续信源的数学模型为

$$X:\begin{bmatrix}\mathbf{R}\\ p(x)\end{bmatrix}$$

并满足

$$\int_{\boldsymbol{R}} p(x)\,\mathrm{d}x = 1$$

其中，$\mathbf{R}$ 是实数域，表示 X 的取值范围。

对于取值范围有限的连续信源，其数学模型还可表示成

$$X:\begin{bmatrix}(a,b)\\ p(x)\end{bmatrix}$$

并满足

$$\int_a^b p(x)\,\mathrm{d}x = 1$$

其中，(a,b) 是 X 的取值范围。

通过对连续变量的取值进行量化分层，可以将连续随机变量用离散随机变量来逼近。量化间隔越小，离散随机变量与连续随机变量越接近。当量化间隔趋于 0 时，离散随机变量就变成了连续随机变量。通过对离散随机变量的熵取极限，可以推导出连续随机变量熵的计算公式。

假定概率密度函数 $p(x)$ 如图 2.5 所示，把连续随机变量 X 的取值分割成 n 个小区间，各小区间等宽，区间宽度 $\Delta=\dfrac{b-a}{n}$，则变量落在第 i 个小区间的概率为

$$P[a+(i-1)\Delta \leqslant x \leqslant a+i\Delta] = \int_{a+(i-1)\Delta}^{a+i\Delta} p(x)\,\mathrm{d}x = p(x_i)\Delta \tag{2.25}$$

其中，x_i 是 $a+(i-1)\Delta$ 到 $a+i\Delta$ 之间的某个值。当 $p(x)$ 是 X 的连续函数时，由中值定理可知，必存在一个 x_i 值使上式成立，这样，连续变量 X 就可用取值为 $x_i(i=1,2,\cdots,n)$ 的离散变量来近似，连续信源就被量化成离散信源。

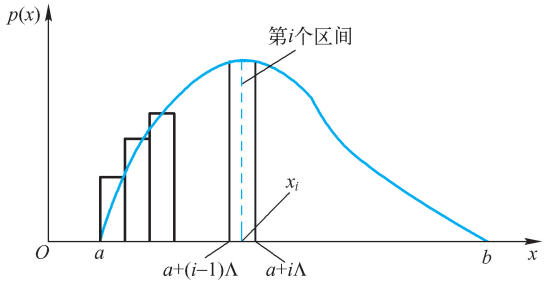


图 2.5　连续随机变量的概率密度函数

这 n 个取值对应的概率分布为 $p_i=p(x_i)\Delta$，这时的离散信源熵是

$$\begin{aligned}H(X) &= -\sum_{i=1}^{n} p(x_i)\Delta\log[p(x_i)\Delta]\\ &= -\sum_{i=1}^{n} p(x_i)\Delta\log p(x_i) - \sum_{i=1}^{n} p(x_i)\Delta\log\Delta\end{aligned} \tag{2.26}$$

当 $n\to\infty$，$\Delta\to 0$ 时，如果上式极限存在，离散信源熵就变成了连续信源的熵：

$$\lim_{\substack{n\to\infty\\ \Delta\to 0}} H(X) = \lim_{\substack{n\to\infty\\ \Delta\to 0}} -\sum_{i=1}^{n} p(x_i)\Delta \log p(x_i) - \lim_{\substack{n\to\infty\\ \Delta\to 0}} \sum_{i=1}^{n} p(x_i)\Delta \log \Delta \tag{2.27}$$

$$= -\int_a^b p(x)\log p(x)\,\mathrm{d}x - \lim_{\substack{n\to\infty\\ \Delta\to 0}} \log\Delta \int_a^b p(x)\,\mathrm{d}x \tag{2.28}$$

$$= -\int_a^b p(x)\log p(x)\,\mathrm{d}x - \lim_{\substack{n\to\infty\\ \Delta\to 0}} \log\Delta \tag{2.29}$$

式(2.29)中的第一项一般是定值，第二项为无穷大量，因此连续信源的熵实际上是无穷大量。这一点是可以理解的，因为连续信源的可能取值是无限多的，它的不确定性是无限大的，当确知输出为某值后，获得的信息量也是无限大的。在丢掉第二项后，我们定义第一项为连续信源的微分熵：

$$h(X) = -\int_{\mathbf{R}} p(x)\log p(x)\,\mathrm{d}x \tag{2.30}$$

微分熵又称为差熵。虽然 $h(X)$ 已不能代表连续信源的平均不确定性，也不能代表连续信源输出的信息量，但是它具有与离散熵相同的形式，也满足离散熵的主要特性，如可加性，但是不具有非负性。另外，我们在实际问题中常常考虑的是熵差，如平均互信息，在讨论熵差时，只要两者离散逼近时所取的间隔 Δ 一致，这两个无限大量将互相抵消，所以熵差具有信息的特性，如非负性。由此可见，连续信源的熵 $h(X)$ 具有相对性。

同样，可以定义两个连续随机变量的联合熵为

$$h(XY) = -\iint_{\mathbf{R}^2} p(xy)\log p(xy)\,\mathrm{d}x\mathrm{d}y \tag{2.31}$$

及条件熵为

$$h(Y\mid X) = -\iint_{\mathbf{R}^2} p(xy)\log p(y\mid x)\,\mathrm{d}x\mathrm{d}y \tag{2.32}$$

$$h(X\mid Y) = -\iint_{\mathbf{R}^2} p(xy)\log p(x\mid y)\,\mathrm{d}x\mathrm{d}y \tag{2.33}$$

并且它们之间也有与离散随机变量一样的相互关系，即

$$h(XY) = h(X) + h(Y\mid X) = H_C(Y) + h(X\mid Y) \tag{2.34}$$

$$h(X\mid Y) \leqslant h(X) \tag{2.35}$$

$$h(Y\mid X) \leqslant H_C(Y) \tag{2.36}$$

【例 2.8】X 是在区间(a,b)内服从均匀分布的连续随机变量，求微分熵。

$$p(x) = \begin{cases} \dfrac{1}{b-a} & x\in(a,b) \\ 0 & x\notin(a,b) \end{cases}$$

【解】

$$h(X) = -\int_a^b p(x)\log p(x)\,\mathrm{d}x$$

$$= -\int_a^b \frac{1}{b-a}\log\frac{1}{b-a}\mathrm{d}x$$

$$= \log(b-a)$$

当$(b-a)>1$ 时，$h(X)>0$；

当$(b-a)=1$时，$h(X)=0$；

当$(b-a)<1$时，$h(X)<0$。

这说明连续熵不具有非负性，失去了信息的部分含义和性质，但是熵差具有信息的特性。

【例 2.9】求均值为 m、方差为 σ^2 的高斯分布的随机变量的微分熵。

【解】高斯随机变量的概率密度为

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-m)^2}{2\sigma^2}}$$

微分熵

$$\begin{aligned}
h(X) &= -\int_{-\infty}^{+\infty}p(x)\log p(x)\,\mathrm{d}x \\
&= -\int_{-\infty}^{+\infty}p(x)\log\left[\frac{1}{\sqrt{2\pi\sigma}}\mathrm{e}^{\frac{(x-m)^2}{2\sigma^2}}\right]\mathrm{d}x \\
&= -\int_{-\infty}^{+\infty}p(x)\log\frac{1}{\sqrt{2\pi}\sigma}-\int_{-\infty}^{+\infty}p(x)\left[-\frac{(x-m)^2}{2\sigma^2}\right]\mathrm{d}x\log\mathrm{e} \\
&= \log\sqrt{2\pi\sigma}+\int_{-\infty}^{+\infty}p(x)\frac{(x-m)^2}{2\sigma^2}p(x)\,\mathrm{d}x\log\mathrm{e} \\
&= \log\sqrt{2\pi}\sigma+\frac{1}{2}\log\mathrm{e} \\
&= \log\sqrt{2\pi\mathrm{e}}\sigma
\end{aligned}$$

这里对数以 2 为底，微分熵的单位为比特/样值，如果对数取以 e 为底，则得到

$$h(X)=\ln\sqrt{2\pi\mathrm{e}}\sigma \text{ 奈特/样值}$$

可以看到，正态分布的连续信源的微分熵与数学期望 m 无关，只与方差 σ^2 有关。

【例 2.10】求指数分布的随机变量的微分熵。

$$p(x)=\begin{cases}\dfrac{1}{a}\mathrm{e}^{-\frac{x}{a}} & x>0\\ 0 & x\leqslant 0\end{cases}$$

【解】

$$\begin{aligned}
h(X) &= -\int_{-\infty}^{+\infty}p(x)\ln p(x)\,\mathrm{d}x \\
&= -\int_{0}^{+\infty}p(x)\ln\left(\frac{1}{a}\mathrm{e}^{-\frac{x}{a}}\right)\mathrm{d}x \\
&= -\int_{0}^{+\infty}p(x)\ln\frac{1}{a}\mathrm{d}x-\int_{0}^{+\infty}p(x)\ln\mathrm{e}^{-\frac{x}{a}}\mathrm{d}x \\
&= \ln a\int_{0}^{+\infty}p(x)\,\mathrm{d}x+\frac{1}{a}\ln\mathrm{e}\int_{0}^{+\infty}xp(x)\,\mathrm{d}x \quad \left(\int_{0}^{+\infty}xp(x)\,\mathrm{d}x=a,\int_{0}^{+\infty}p(x)\,\mathrm{d}x=1\right) \\
&= \ln a+\ln\mathrm{e}=\ln a\mathrm{e}
\end{aligned}$$

所以，指数分布的相对熵只取决于信源的均值 a。

【例 2.11】求 N 维高斯信源的熵。

【解】

把 N 维高斯信源输出的 N 维连续随机矢量记为列向量，则其转置为行向量

$$\boldsymbol{X}=[X_1,X_2,\cdots,X_N]^{\mathrm{T}}$$

其均值矢量为

$$\boldsymbol{M}=[m_1,m_2,\cdots,m_N]^{\mathrm{T}}$$

协方差矩阵为

$$\boldsymbol{R}=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1N} \\ r_{21} & r_{22} & \cdots & r_{2N} \\ \vdots & \vdots & & \vdots \\ r_{N1} & r_{N2} & \cdots & r_{NN} \end{bmatrix}$$

其中

$$r_{ij}=E[(x_i-m_i)(x_j-m_j)]\,(i,j=1,2,\cdots,N)$$

N 维联合概率密度为

$$p(x_1x_2\cdots x_N)=\frac{1}{(2\pi)^{\frac{N}{2}}|\boldsymbol{R}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{M})^{\mathrm{T}}\boldsymbol{R}^{-1}(\boldsymbol{X}-\boldsymbol{M})\right\}$$

N 维联合熵为

$$\begin{aligned} &h(X_1X_2\cdots X_N)\\ &=-\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}p(x_1x_2\cdots x_N)\ln p(x_1x_2\cdots x_N)\,\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_N\\ &=-\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}p(x_1x_2\cdots x_N)\left[-\ln\sqrt{(2\pi)^N|\boldsymbol{R}|}-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{M})^{\mathrm{T}}\boldsymbol{R}^{-1}(\boldsymbol{X}-\boldsymbol{M})\right]\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_N\\ &=\frac{1}{2}\ln[(2\pi)^N|\boldsymbol{R}|]+\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\frac{1}{2}(\boldsymbol{X}-\boldsymbol{M})^{\mathrm{T}}\boldsymbol{R}^{-1}(\boldsymbol{X}-\boldsymbol{M})p(x_1x_2\cdots x_N)\,\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_N\\ &=\frac{1}{2}\ln[(2\pi)^N|\boldsymbol{R}|]+\frac{N}{2} \end{aligned}$$

当 $X_1,X_2,\cdots,X_N$ 统计独立时，$|\boldsymbol{R}|=\prod_{i=1}^{N}\sigma_i^2$，这时有

$$h(X_1X_2\cdots X_N)=\frac{1}{2}\sum_{i=1}^{N}\ln\sigma_i^2+\frac{N}{2}\ln 2\pi+\frac{N}{2}$$

2.4.1 连续信源的最大熵

离散信源当信源符号为等概分布时有最大熵。连续信源微分熵也有极大值，但是与约束条件有关，当约束条件不同时，信源的最大熵不同。我们一般关心的是下面两种约束下的最大熵。

【定理 2-2】在输出幅度受限的情况，服从均匀分布的随机变量 X 具有最大输出熵。

【证明】设输出幅度限制在 $[a,b]$ 内，则约束条件为

$$\int_a^b p(x)\,\mathrm{d}x=1$$

因此，这是在约束条件下求极值的问题，用拉格朗日乘子法。

令

$$F[p(x)]=h(X)+\lambda\int_a^b p(x)\,\mathrm{d}x$$

由$\frac{\partial F}{\partial p(x)}=-\log p(x)-1+\lambda=0$，得$p(x)=e^{\lambda-1}$。

由$\int_a^b p(x)\mathrm{d}x=\int_a^b e^{\lambda-1}\mathrm{d}x=1$，得$e^{\lambda-1}=\frac{1}{b-a}$，即

$$p(x)=\begin{cases}\frac{1}{b-a} & a\leqslant x\leqslant b\\ 0 & \text{其他}\end{cases}$$

从而

$$\begin{aligned}h(X)&=-\int_a^b p(x)\log p(x)\mathrm{d}x\\&=-\int_a^b\frac{1}{b-a}\log\frac{1}{b-a}\mathrm{d}x\\&=\log(b-a)\end{aligned}$$

因此，对于输出信号幅度受限的连续信源，当满足均匀分布时达到最大熵。这个结论与离散信源在等概分布时达到最大熵的结论类似。

【定理 2-3】对于平均功率受限的连续随机变量，当服从高斯分布时具有最大熵。

【证明】对于均值为m、方差为σ^2的连续随机变量，平均功率P=直流功率+交流功率$=m^2+\sigma^2$。因此，平均功率受限相当于约束条件

$$\int_{-\infty}^{+\infty}p(x)\mathrm{d}x=1$$

$$\int_{-\infty}^{+\infty}xp(x)\mathrm{d}x=m$$

$$\int_{-\infty}^{+\infty}(x-m)^2p(x)\mathrm{d}x=\sigma^2$$

这仍然是在约束条件下求极值的问题。令

$$F[p(x)]=h(X)+\lambda_1\int_{-\infty}^{+\infty}p(x)\mathrm{d}x+\lambda_2\int_{-\infty}^{+\infty}xp(x)\mathrm{d}x+\lambda_3\int_{-\infty}^{+\infty}(x-m)^2p(x)\mathrm{d}x$$

由$\frac{\partial F[p(x)]}{\partial p(x)}=-\ln p(x)-1+\lambda_1+\lambda_2x+\lambda_3(x-m)^2=0$，得$p(x)=e^{\lambda_3(x-m)^2+\lambda_2x+\lambda_1-1}$。代入约束条件中，可得

$$e^{\lambda_1-1}=\frac{1}{\sqrt{2\pi}\sigma}$$

$$\lambda_2=0$$

$$\lambda_3=-\frac{1}{2\sigma^2}$$

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\}$$

可以求出$h(X)=\log\sqrt{2\pi e}\sigma$（见例 2.9）。

这说明当平均功率受限时，高斯分布的连续信源的熵最大，也就是说，高斯信源输出的每个样值（也称为自由度）提供的平均信息量最大，其大小随交流功率σ^2的增加而增加。

2.4.2 连续信源的熵功率

与离散信源一样，在讨论了连续信源的最大熵问题之后，我们也要考虑没有达到最大熵

的信源的冗余度问题。从这个角度出发，我们定义了熵功率的概念。在不同的约束条件下连续信源有不同的最大熵，因为均值为零、平均功率受限的连续信源是实际最常见的一种信源，下面重点讨论这种信源的冗余问题。

均值为零、平均功率限定为 P 的连续信源服从高斯分布时达到最大熵：

$$h_0(X)=\ln\sqrt{2\pi e\sigma^2}=\ln\sqrt{2\pi eP} \tag{2.37}$$

也就是说，高斯信源的熵值与 P 有确定的对应关系：

$$P=\frac{1}{2\pi e}e^{2h_0(X)} \tag{2.38}$$

如果另一信源的平均功率也为 P，但不是高斯分布，那么它的熵值一定比高斯信源的熵值小。反之，如果有一个信源与这个高斯信源有相同的熵，则它的平均功率 $P'\geqslant\overline{P}$，$\overline{P}$为高斯信源的平均功率。对于非高斯信源，$h(X)\leqslant\ln\sqrt{2\pi eP}$；而对于高斯信源，$h(X)=\ln\sqrt{2\pi eP}$。

现在假定限定的平均功率为 P，某连续信源的熵为 $h(X)$，则与它具有相同熵的高斯信源的平均功率$\overline{P}$定义为熵功率，即

$$\overline{P}=\frac{1}{2\pi e}e^{2h(X)} \tag{2.39}$$

从而

$$\overline{P}\leqslant P$$

当该连续信源为高斯信源时，等号成立。

$\overline{P}$可以表示连续信源剩余的大小。如果熵功率等于信源平均功率，表示信源没有剩余；熵功率与信源的平均功率相差越大，说明信源的剩余越大。所以，信源平均功率和熵功率之差 $P-\overline{P}$被称为连续信源的剩余度。

扩展阅读：随机过程

如果每次随机试验的结果是一个时间 t 的函数，这样的随机现象称为**随机过程**$\{X(e,t),t\in T\}$。比如热噪声电压，每次观测的结果都不同。如果固定时间 t，则 $X(e,t_j)$ 就是一个定义在样本空间$\{e\}$上的随机变量。

随机过程可依其在任一时刻的随机变量的取值 $X(e,t_j)$ 是连续型还是离散型，分成连续型随机过程（如热噪声电压）和离散型随机过程。比如，连续抛掷一颗骰子的试验，设 X_n 是第 n 次$(n\geqslant1)$抛掷的点数，对于 $n=1,2,\cdots$的不同值，X_n 是不同离散取值的随机变量，因而$\{X_n,n\geqslant1\}$构成一离散型随机过程。时间参数的取值是离散的随机过程（如掷骰子试验），称为随机序列。

随机过程的统计特性通常用一维和多维随机变量的统计特性来描述。

在随机过程中，我们主要研究平稳随机过程，它的统计特性不随时间的推移而变化，也就是说，平稳随机过程的任意 n 维随机变量的概率分布与时间起点无关，任意 n 维随机变量具有相同的概率分布。

对任意 $n(n=1,2,\cdots)$，$t_1,t_2,\cdots,t_n\in T$ 和任意实数 $h,t_1+h,t_2+h,\cdots(t_n+h\in T)$，$n$ 维随机变量 $(X(t_1),X(t_2),\cdots,X(t_n))$ 和 $(X(t_1+h),X(t_2+h),\cdots,X(t_n+h))$ 具有相同的分布函数。

参数 t 为离散取值的平稳随机过程又称为平稳随机序列。

对于非平稳随机过程，我们主要研究的是马尔可夫过程，它满足以下统计关系：

$$\begin{aligned}&P\{X(t_{n+1})\leqslant x_{n+1}\mid X(t_1)=x_1,X(t_2)=x_2,\cdots,X(t_n)=x_n\}\\&=P\{X(t_{n+1})\leqslant x_{n+1}\mid X(t_n)=x_n\}\quad(x_n\in R)\end{aligned}$$

也就是在已知前面 n 个时刻的状态的条件下，第 $n+1$ 时刻的条件分布函数恰等于在已知第 n 时刻的状态的条件下第 $n+1$ 时刻的条件分布函数。马尔可夫过程的将来与现在有关，与过去无关，将来通过现在与过去发生联系。马尔可夫过程在满足某些条件成为遍历的马尔可夫过程后是一个平稳随机过程。

状态、时间都是离散的马尔可夫过程称为马尔可夫链。

马尔可夫链通常用它的转移概率来表示：$p_{ij}(m,m+n)=P\{X_{m+n}=s_j\mid X_m=s_i\}$，它表示马尔可夫链在时刻 m 处于状态 s_i 的条件下，在时刻 $m+n$ 转移到状态 s_j 的转移概率。马尔可夫链在时刻 m 可以处于该时刻的任一状态，到另一时刻 $m+n$ 也可以转移到该时刻的任一状态。因此，由转移概率可以组成马尔可夫链的转移概率矩阵：

$$\boldsymbol{P}(m,m+n)=(p_{ij}(m,m+n))$$

由于马尔可夫链在时刻 m 从任一状态 s_i 出发，到另一时刻 $m+n$ 必然转移到状态集 $S=\{s_1,s_2,\cdots s_i,\cdots s_J\}$ 中的某个状态，所以

$$\sum_{j=1}^{\infty}p_{ij}(m,m+n)=1\qquad i=1,2,\cdots$$

并且 $p_{ij}(m,n)\geqslant 0$，$s_i,s_j\in S$。

当 $n=1$ 时，$p_{ij}(m,n)$ 记为 $p_{ij}(m)$，$m\geqslant 0$，称为基本转移概率，也称为一步转移概率：

$$\begin{aligned}p_{ij}^{(1)}(m)&=p_{ij}(m)\\&=P\{X_{m+1}=s_j\mid X_m=s_i\}\end{aligned}$$

把 k 步转移概率写成

$$p_{ij}^{(k)}(m)=P\{X_{m+k}=s_j\mid X_m=s_i\}$$

它表示在时刻 m，X_m 的状态为 s_i 的条件下，经过 k 步转移到达状态 s_j 的概率。

转移概率矩阵 $\boldsymbol{P}^{(k)}(m)=\{p_{ij}^{(k)}(m),s_i,s_j\in S\}$ 称为 k 步转移矩阵。一步转移概率矩阵为 $\boldsymbol{P}(m)=\{p_{ij}(m),s_i,s_j\in S\}$。

当转移概率 $p_{ij}(m,m+n)$ 只与状态 s_i、s_j 及时间间距 n 有关，而与时间起点无关，即 $p_{ij}(m,m+n)=p_{ij}(n)$ 时，则称这类马尔可夫链为时齐马尔可夫链或齐次马尔可夫链，也称为具有平稳转移概率的马尔可夫链。这里的平稳仅仅是转移概率的平稳，还不是平稳过程。

由一步转移概率 p_{ij} 可以写出其一步转移概率矩阵：

$$\boldsymbol{P}=\{p_{ij},s_i,s_j\in S\}$$

或

$$\boldsymbol{P}=\begin{bmatrix}p_{11}&p_{12}&p_{13}&\cdots\\p_{21}&p_{22}&p_{23}&\cdots\\p_{31}&p_{32}&p_{33}&\cdots\\\cdots&\cdots&\cdots&\cdots\end{bmatrix}$$

矩阵中的每个元素都是非负的，并且每行之和均为 1。如果马尔可夫链中的状态空间 $S=\{s_1,s_2,\cdots,s_J\}$ 是有限的，则称为有限状态的马尔可夫链。如果状态空间 $S=\{s_1,s_2,\cdots\}$ 是无穷集合，则称为可数无穷状态的马尔可夫链。

对于具有 $m+n$ 步转移概率的齐次马尔可夫链，存在下述 C－K 方程：

$$\boldsymbol{P}^{(m+n)}=\boldsymbol{P}^{(m)}\boldsymbol{P}^{(n)}$$

或

$$p_{ij}^{(m+n)}=\sum_k p_{ik}^{(m)}p_{kj}^{(n)}$$

对于齐次马尔可夫链来说，一步转移完全决定了 k 步转移概率 $\boldsymbol{P}^{(k)}=(\boldsymbol{P})^k$。

下面求在某时刻的状态的概率分布 $P(X_k=s_j)$。记 $p_{0i}=P(X_0=s_i)$ 表示初始概率，则

$$\begin{aligned}P(X_k=s_j)&=\sum_i P(X_k=s_j,X_0=s_i)\\&=\sum_i p_{0i}p_{ij}^{(k)}\end{aligned}$$

一般情况下，绝对概率与初始分布 p_{0i} 有关，但当 $\lim\limits_{k\to\infty}p_{ij}^{(k)}$ 极限存在且等于一个与起始状态 s_i 无关的被称为平稳分布的 W_j，即 $\lim\limits_{k\to\infty}p_{ij}^{(k)}=W_j$ 与 s_i 无关时，则不论起始状态是什么，此马尔可夫链最终可以达到稳定，稳定后所有时刻随机变量的概率分布不变。这时，有

$$\begin{aligned}\lim_{k\to\infty}P(X_k=s_j)&=\lim_{k\to\infty}\sum_i p_{0i}p_{ij}^{(k)}\\&=\sum_i p_{0i}W_j\\&=W_j\end{aligned}$$

马尔可夫链达到了稳定的分布，稳定分布只与转移概率有关，而与初始分布无关。因此，经过长时间的转移，到达某状态时的概率与初始状态无关，起始状态只使前面有限个变量的分布改变，如同电路中的暂态和稳态一样。

所以，若齐次马尔可夫链对一切 i 和 j 存在不依赖于 i 的极限：

$$\lim_{k\to\infty}p_{ij}^{(k)}=W_j\text{ 且 }W_j=\sum_i W_i p_{ij}\quad\left(\sum_j W_j=1,W_j\geqslant 0\right)$$

则称其具有遍历性（各态历经性），这时的分布 W_j 称为极限分布或平稳分布。

遍历性的直观意义是，不论起始状态是哪个状态 s_i，当转移步数 k 足够大时，转移到状态 s_j 的概率 $p_{ij}^{(k)}$ 都近似等于某个常数 W_j；反之，如果转移步数 k 充分大，就可以用常数 W_j 作为 k 步转移概率 $p_{ij}^{(k)}$ 的近似值。这意味着，马尔可夫信源在初始时刻可以处在任意状态，然后状态之间可以任意转移，经过足够长的时间之后，信源处在什么状态已与初始状态无关。这时，每种状态出现的概率已达到一种稳定分布，即平稳分布。就像电路经过暂态后进入稳态一样，到这时，信源才是一个离散平稳信源。

【定理 2-4】 W_j 是满足方程组 $\boldsymbol{WP}=\boldsymbol{W}$ 和 $\sum\limits_j W_j=1$（$W_j\geqslant 0$）的唯一解。

事实上，用 $\lim\limits_{k\to\infty}p_{ij}^{(k)}=W_j$ 求稳态分布是比较困难的。如果能判断稳态分布存在，一般用解方程组 $W_j=\sum\limits_i W_i p_{ij}$，$\sum\limits_j W_j=1$（$W_j\geqslant 0$）来求 W_j。怎样判断马尔可夫链的稳态分布存在呢？

【定理 2-5】 设 $\boldsymbol{P}$ 为某一马尔可夫链的状态转移矩阵，则该马尔可夫链稳态分布存在的

充要条件是，存在一个正整数 N，使矩阵 $\boldsymbol{P}^N$ 中的所有元素均大于零。

【例 2.12】设有一马尔可夫链，其状态转移矩阵为

$$\boldsymbol{P}=\begin{bmatrix}0 & 0 & 1\\ \frac{1}{2} & \frac{1}{3} & \frac{1}{6}\\ \frac{1}{2} & \frac{1}{2} & 0\end{bmatrix}$$

是否存在稳态分布？如果存在，求其稳态分布。

【解】

为了验证它是否满足定理 2-5 的条件，下面计算矩阵

$$\boldsymbol{P}^2=\begin{bmatrix}0 & 0 & *\\ * & * & *\\ * & * & 0\end{bmatrix}\begin{bmatrix}0 & 0 & *\\ * & * & *\\ * & * & 0\end{bmatrix}=\begin{bmatrix}* & * & 0\\ * & * & *\\ * & * & *\end{bmatrix}$$

$$\boldsymbol{P}^3=\begin{bmatrix}* & * & 0\\ * & * & *\\ * & * & *\end{bmatrix}\begin{bmatrix}0 & 0 & *\\ * & * & *\\ * & * & 0\end{bmatrix}=\begin{bmatrix}* & * & *\\ * & * & *\\ * & * & *\end{bmatrix}$$

其中，$*$ 表示非零元素。因此，这个马尔可夫链是遍历的，其稳态分布存在。

由定理 2-4，$\boldsymbol{W}$ 满足方程组 $\boldsymbol{WP}=\boldsymbol{W}$ 且 $\sum_j W_j=1$ $(W_j\geqslant 0)$，将矢量 $\boldsymbol{W}$ 写成分量的形式 $\boldsymbol{W}=(W_1\ W_2\ W_3)$，代入 $\boldsymbol{WP}=\boldsymbol{W}$，得到

$$\begin{cases}\frac{1}{2}W_2+\frac{1}{2}W_3=W_1\\ \frac{1}{3}W_2+\frac{1}{2}W_3=W_2\\ W_1+\frac{1}{6}W_2=W_3\\ W_1+W_2+W_3=1\end{cases}$$

求得 $W_1=\frac{1}{3}$，$W_2=\frac{2}{7}$，$W_3=\frac{8}{21}$。

定理 2-4 给定的条件等价于存在一个状态 s_j 和正整数 N，使得从任意原始状态出发，经过 N 步转移之后，一定可以到达状态 s_j。也就是说，只有在转移一定步数后，各状态之间均可相通的条件下，当转移步数足够大，各状态出现的概率才能稳定在某一极限值，存在状态的极限概率。所谓“各态历经”，其含义之一就是各态相通，均可经历；其含义之二就是由各态历经过程产生的每个序列，都有相同的统计特性。如果 $\boldsymbol{P}$ 不含零元素，即任一状态经一步转移便可达其他状态，则稳态分布必然存在。

时齐马尔可夫链可以用状态转移图来表示，可以判断状态相通的情况。图 2.6 是一个有着 6 个状态的时齐马尔可夫链。时齐马尔可夫链的状态可以根据其性质分为常返态和过渡态。常返态是从该状态出发，经过若干步以后总能回到该状态，如图中的状态 2、3、4、5、6 均为常返态。而过渡态是从该状态出发能到达某一个其他状态，但不能从其他状态返回，

如图中的状态1。常返态中又分为周期态和非周期态。周期态指存在某大于1的整数d，当n能被d整除时，$p_{ii}^{(n)}>0$，而不能被d整除的n，则$p_{ii}^{(n)}=0$。如状态4、5，周期$d=2$。非周期的常返态称为遍历状态，如状态2、3。

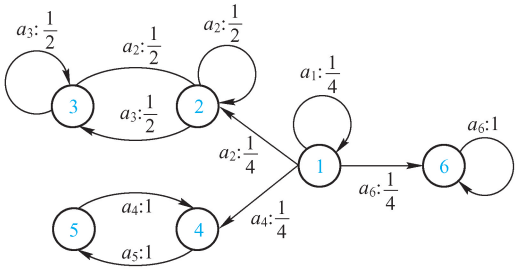


图2.6　状态转移图

若状态空间的某一子集中的任何一状态都不能到达子集以外的任何状态，则称该子集为闭集，如{2,3,4,5}，{2,3}，{4,5}，{6}。闭集中不包含其他非空闭集，称为不可约的(不可分的，既约的)，如{2,3}，{4,5}，{6}。

从不可约的非周期态出发，在转移一定步数后，各状态之间可相通，经过足够长时间后，就可以使各状态出现的概率稳定在某一极限值。

既约、非周期、有限状态的马尔可夫链，其n步转移概率在n很大时趋于一个和初始状态无关的极限概率W_j，它是满足方程组$W_j=\sum_i W_i p_{ij}$和$\sum_j W_j=1$（$W_j\geqslant 0$）的唯一解，称W_j为马尔可夫链的平稳分布，它是当时间足够长之后系统处于状态s_j的概率，此时马尔可夫链是平稳的或称为遍历的。

有了马尔可夫链的这些知识，我们可以求马尔可夫信源的极限熵。

扩展阅读：隐马尔可夫模型与赌场风云

隐马尔可夫模型（Hidden Markov Model，HMM）作为一种统计分析模型，创立于20世纪70年代。自20世纪80年代以来，HMM被应用于语音识别，并取得了重大成功。到了90年代，HMM还被引入计算机文字识别和移动通信核心技术“多用户的检测”。HMM在生物信息科学、故障诊断等领域也开始得到应用。

在简单的马尔可夫模型（如马尔可夫链）中，状态是直接可见的，因此状态转移概率是唯一的参数。在隐马尔可夫模型中，状态是不直接可见的，但依赖于该状态下的输出是可见的。怎么通过可见的输出返回去求状态的转移概率是HMM的研究内容。怎么理解呢，还是用最经典的例子，掷骰子。假设我手里有三颗不同的骰子。第一颗骰子是我们平常见的骰子（称这颗骰子为D6），6个面，每个面（1,2,3,4,5,6）出现的概率是$\frac{1}{6}$。第二颗骰子是个四面体（称这个骰子为D4），每个面（1,2,3,4）出现的概率是$\frac{1}{4}$。第三颗骰子有8个面（称这个骰子为D8），每个面（1,2,3,4,5,6,7,8）出现的概率是$\frac{1}{8}$。

假设开始掷骰子，先从三颗骰子里挑一颗，挑到每颗骰子的概率都是$\frac{1}{3}$。然后掷骰子，得到一个数字，即1、2、3、4、5、6、7、8中的一个。

不停地重复上述过程，我们会得到一串数字，每个数字都是1~8中的一个。例如，可能得到这么一串数字（掷骰子10次）：1 6 3 5 2 7 3 5 2 4。这串数字叫做可见量链。但是在隐马尔可夫模型中，不仅有这么一串可见量链，还有一串隐含量链。例如，这串隐含变量链就是你用的骰子的序列。比如，隐含量链有可能是D6 D8 D8 D6 D4 D8 D6 D6 D4 D8。

一般来说，HMM中说到的马尔可夫链其实是指隐含量链，因为隐含量（骰子）之间存在转换概率。在本例中，D6的下一个状态是D4、D6、D8的概率都是$\frac{1}{3}$。D4、D8的下一个状态是D4、D6、D8的转换概率也都一样是$\frac{1}{3}$。这样设定是为了最开始容易说清楚，但是其实可以随意设定转换概率或转换概率分布。比如，可以这样定义，D6后面不能接D4，D6后面为D6的概率是0.9，为D8的概率是0.1。这样就是一个新的HMM。

同样，尽管可见量之间没有转换概率，但隐含量和可见量之间有一个概率叫做发射概率（emission probability）。对于本例，六面均匀骰子D6产生1、2、3、4、5、6的发射概率均为$\frac{1}{6}$。

最近一个赌场的老板发现生意不佳，手下报告，有位大叔在赌场中总能赢到钱，玩得一手好骰子，每次开局，骰子飞出，沉稳落地，几乎是战无不胜。老板根据多年的经验，推测这位不善之客使用了“偷换骰子大法”，用兜里自带的骰子偷偷换掉了均匀的骰子。老板思考怎么破了这位大叔的局。这时候有个年轻人进来，告诉他用HMM模型可以很好地解决。只要在远处装个摄像头，把每局的骰子的点数都记录下来，就可以用这些数据推导出：① 该大叔是不是在出千？② 如果是在出千，那么他用了几个作弊的骰子？当前是不是在用作弊的骰子？③ 这几个作弊骰子出现各点的概率是多少？这样，不用近身就能算出是不是在作弊，甚至能算出他作弊的骰子是什么样的。那么，只要他再作弊，派人围捕他，当场验证骰子就能让他哑口无言。老板听完年轻人的讲解，连夸这个青年能学以致用，有出息。

这里，大叔使用骰子的状态，如正常骰子、作弊骰子1、作弊骰子2……，是外界不便观察（或观察不到）的状态，是隐含量链，而大叔掷骰子掷出的点数是可见量链。

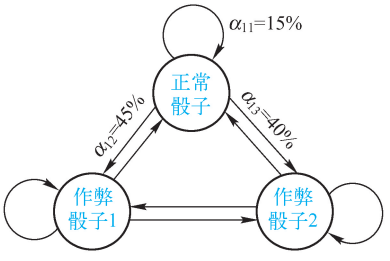


图2.7　隐性状态转移概率
（大叔切换骰子的概率）

HMM模型可以从可见量链分析得到系统隐性状态的转移概率（隐含量链），也就是大叔切换骰子的概率，图2.7是一个例子。

HMM模型也可以分析得到隐含量和可见量之间的发射概率，也就是每个骰子出现各点的概率分布，如图2.8所示。

隐性状态转移概率加上隐含量和可见量之间的发射概率，就是整个HMM模型。总之，HMM模型能够描述扔骰子大叔作弊的频率（骰子更换的概率）和大叔用的骰子的概率分布。

HMM能处理如下三个问题。

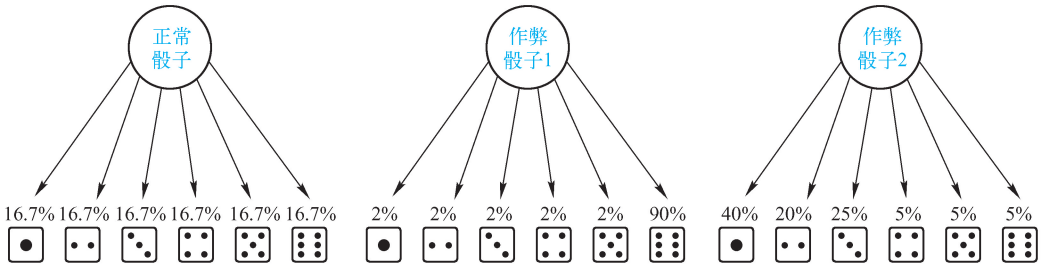


图 2.8　隐含量和可见量之间的发射概率（骰子点数分布概率）

1. 解码（Decoding）

解码就是需要从一连串的骰子结果序列中，看出哪些骰子是用于作弊的骰子，哪些是正常的骰子。图 2.9 给出了一串骰子结果序列(3,6,1,2,…)和大叔的 HMM 模型，我们想要计算哪段骰子结果序列（隐性状态的表现，可见量链）可能对应哪种骰子（隐性状态，隐含量链）。

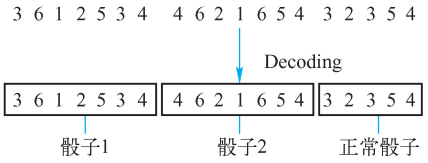


图 2.9　解码

2. 学习（Learning）

学习就是从一连串的骰子结果序列中，学习到大叔切换骰子的概率，当然也有这些骰子的点数的分布概率，得到该大叔的 HMM 模型。这是 HMM 最恐怖也是最复杂的招数！

3. 估计（Evaluation）

估计是指在已经知道该大叔的 HMM 模型的情况下，估测某串骰子出现的可能性。比如，我们能够直接估测到大叔扔到 10 个 6 或 8 个 1 的概率。

至于 HMM 是怎么做到的，感兴趣的同学可以查询相关参考资料。

——根据 ppn029012 的博客改写而成。

习　题　2

2.1　证明 $\lim\limits_{n\to\infty}\frac{1}{2}H(X_nX_{n-1}\mid X_1\cdots X_{n-2})=H_\infty$。

2.2　有一无记忆信源的符号集为{0,1}，已知信源的概率空间为

$$\begin{bmatrix}X\\P\end{bmatrix}=\begin{bmatrix}0 & 1\\ \frac{1}{4} & \frac{3}{4}\end{bmatrix}$$

（1）求信源熵。

（2）求由 m 个“0”和 $100-m$ 个“1”构成的某一特定序列自信息量的表达式。

（3）计算由 100 个符号构成的符号序列的熵。

2.3 有一离散无记忆信源，其输出为 $X\in\{0,1,2\}$，相应的概率为 $p_0=\frac{1}{4},p_1=\frac{1}{4},p_2=\frac{1}{2}$，设计两个独立实验去观察它，其结果分别为 $Y_1\in\{0,1\},Y_2\in\{0,1\}$，已知条件概率如题表 2.3 所列。

题表 2.3

$p(y_1\mid x)$	0	1	$p(y_2\mid x)$	0	1
0	1	0	0	1	0
1	0	1	1	1	0
2	$\frac{1}{2}$	$\frac{1}{2}$	2	0	1

（1）求 $I(X;Y_1)$ 和 $I(X;Y_2)$，并判断哪一个实验好一些。

（2）求 $I(X;Y_1Y_2)$，并计算做 Y_1 和 Y_2 两个实验比做 Y_1 或 Y_2 中的一个实验可多得多少关于 X 的信息。

（3）求 $I(X;Y_1\mid Y_2)$ 和 $I(X;Y_2\mid Y_1)$，并解释它们的含义。

2.4 某信源符号集的概率分布和对应的二进制代码如题表 2.4 所示。

题表 2.4

信源符号	u_0	u_1	u_2	u_3
概率	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$
代码	0	10	110	111

（1）求信源符号熵。

（2）求平均每个消息符号所需要的二进制码元的个数或平均代码长度。进而用这一结果求码序列中的二进制码元的熵。

（3）当消息由符号序列组成时，各符号之间若相互独立，求其对应的二进码序列中出现“0”和“1”的无条件概率 $p(0)$ 和 $p(1)$，以及相邻码元间的条件概率 $p(0\mid 0)$、$p(1\mid 0)$、$p(0\mid 1)$ 和 $p(1\mid 1)$。

2.5 二次扩展信源的熵为 $H(X^2)$，而一阶马尔可夫信源的熵为 $H(X_2\mid X_1)$。试比较两者的大小，并说明原因。

2.6 一个马尔可夫过程的基本符号为 0、1、2，这三个符号等概出现，并且具有相同的转移概率。

（1）画出一阶马尔可夫过程的状态图，并求稳定状态下的一阶马尔可夫信源熵 H_1 和信源剩余度。

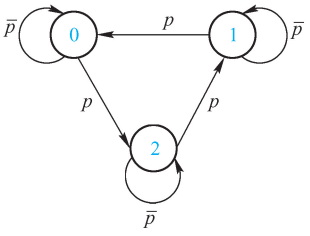
（2）画出二阶马尔可夫过程的状态图，并求稳定状态下二阶马尔可夫信源熵 H_2 和信源剩余度。

2.7 一阶马尔可夫信源的状态转移图如题图 2.7 所示，信源 X 的符号集为 $\{0,1,2\}$。

（1）求平稳后的信源的概率分布。

（2）求信源熵 H_∞。

（3）求当 $p=0$ 或 $p=1$ 时信源的熵，并说明其理由。



题图 2.7

2.8 有一个二元无记忆信源，其发 0 的概率为 p，而 $p\approx 1$，所以在发出的二元序列中经常出现的是那些一串为 0 的序列，称高概率序列。对于这样的信源我们可以用另一新信源来代替，新信源中只包含这些高概率序列。这时新信源 $S_n=\{s_1,s_2,s_3,\cdots,s_n,s_{n+1}\}$，共有 $n+1$ 个符号，它与高概率的二元序列的对应关系如下：

二元序列：$1,01,001,\cdots,00\cdots01$（共 $n-1$ 个 0），$00\cdots000$（共 n 个 0）

新信源符号：$s_1,s_2,s_3,\cdots,s_n,s_{n+1}$

（1）求 $H(S_n)$。

（2）当 $n\to\infty$ 时，求信源的熵 $H(S)=\lim\limits_{n\to\infty}H(S_n)$。

2.9 给定状态转移概率矩阵 $\boldsymbol{P}=\begin{bmatrix}1-p & p\\ 1 & 0\end{bmatrix}$，求：

（1）此两状态马尔可夫链的熵率 H_∞。

（2）此熵率的极大值及相应的 p。

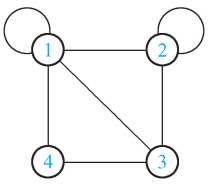
（3）在达到最大熵率的情况下，求出每个 n 长序列的概率。

2.10 在一个 3×3 的国际象棋棋盘上，分别计算“王”、“车”、“左象”、“右象”和“后”随机行走的熵率。

2.11 题图 2.11 是一幅有 4 个节点的随机行走图，从任何一个节点走到下个节点的概率都相等。

（1）求随机行走的稳态分布。

（2）求随机行走的熵率。



题图 2.11

2.12 求具有如下概率密度函数的随机变量的熵。

（1）指数分布 $f(x)=\lambda e^{-\lambda x}$（$x\geqslant 0$）。

（2）$f(x)=\dfrac{1}{2}\lambda e^{-\lambda|x|}$。

（3）单边高斯分布 $f(x)=\dfrac{2}{\sqrt{2\pi\sigma^2}}\lambda e^{-x^2/2\sigma_2}$（$x\geqslant 0$）。

2.13 连续随机变量 X 和 Y 的联合概率密度为

$$p(xy)=\frac{1}{2\pi\sqrt{SN}}\exp\left\{-\frac{1}{2N}\left[x^2\left(1+\frac{N}{S}\right)-2xy+y^2\right]\right\}$$

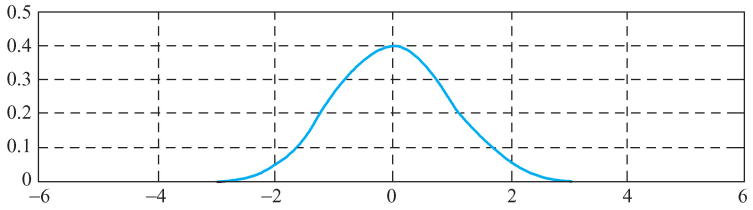
试求 $h(X),h(Y),h(Y|X)$ 和 $I(X;Y)$。

2.14 一信源产生的时不变波形信号（即信号统计特性不随时间而变）的带宽 $W=4\text{kHz}$，幅度分布为

$$p(x)=e^{-x}(x\geqslant 0)$$

设在信号幅度 0～2 区间按量化单位 $\Delta=0.5$ 做量化，试求该信源的信息输出率。

2.15 随机变量 X 和 Y 的联合概率密度函数在曲线 $y=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ 和 X 轴所组成的区域内均匀分布，如题图 2.15 所示。



题图 2.15

（1）求 $h(XY)$。

（2）求 $h(X)$。

（3）Y 的概率密度函数为 $f(y)=2\sqrt{-2\ln(y\sqrt{2\pi})}$，$0<y\leqslant 1/\sqrt{2\pi}$。证明：

$$-\frac{1}{2}\ln 2\pi-\frac{1}{2}<h(Y)<-\frac{1}{2}\ln 2\pi$$

2.16 给定状态转移概率矩阵 $\boldsymbol{P}=\begin{bmatrix}1-\alpha & \alpha\\ \beta & 1-\beta\end{bmatrix}$，求此二状态马尔可夫链的熵率 H_∞。

2.17 布袋中有手感完全相同的 3 个红球和 3 个蓝球，每次从中随机取出一个球，取出后不放回布袋。用 X_i 表示第 i 次（$i=1,2,\cdots,6$）取出的球的颜色。求：

（1）$H(X_1)$。

（2）$H(X_2)$。

（3）$H(X_2\mid X_1)$。

（4）随着 k 的增加，$H(X_k\mid X_1\cdots X_{k-1})$ 是增加还是减少？请解释。

（说明：所有答案用 $H(p)$ 的形式表示。）

2.18 已知一个二元一阶马尔可夫信源的状态转移概率矩阵为 $\boldsymbol{P}=\begin{bmatrix}0.9 & 0.1\\ 0.2 & 0.8\end{bmatrix}$。

（1）求此马尔可夫信源的熵率。

（2）求符号序列 1000011 的概率（根据平稳分布确定第一个符号的概率）。

（3）计算分布函数 $F(\boldsymbol{x})=P_r\{(X_1X_2\cdots)<\boldsymbol{x}\}$ 当 $\boldsymbol{x}=1000011$ 时的值。

2.19 盒子里有两枚偏畸硬币，硬币 1 正面向上的概率为 p，硬币 2 正面向上的概率为 $1-p$，$0<p<0.5$。随机取一枚硬币并且连续投掷。用 $Z\in\{1,2\}$ 表示所选择的硬币，$X_1,X_2,X_3,\cdots$ 表示每次投掷的结果（正面或反面）。

（1）$X_1,X_2,X_3,\cdots$ 是否为平稳过程？是否为马尔可夫过程？

（2）求 $H(X_1X_2\cdots X_n\mid Z)$。

（3）求 $I(X_1;X_2\mid Z)$。

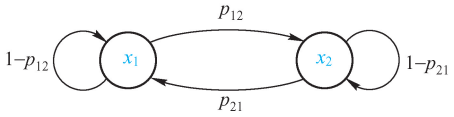
（4）求 $H(X_1X_2)$。

（5）求 $I(X_1;X_2)$ 和 $I(X_3;X_{729})$。

（6）求熵率 $H_\infty=\lim\limits_{n\to\infty}\frac{1}{n}H(X_1X_2\cdots X_n)$。

（7）求熵率$\lim_{n\to\infty}H(Z\mid X_1X_2\cdots X_n)$。

2.20 二元一阶马尔可夫信源的状态转移图如题图 2.20 所示。计算当 $p_{12}=0.2$，$p_{21}=0.3$ 时该马尔可夫信源的熵率，并求具有同样符号概率分布的离散无记忆信源的熵。



题图 2.20

2.21 求具有如下概率密度函数的连续随机变量的微分熵（$\lambda>0$）。

$$fx(x)=\begin{cases}(x+\lambda)/\lambda_2 & -\lambda\leqslant x\leqslant 0\\(-x+\lambda)/\lambda_2 & 0<x\leqslant\lambda\\0 & \text{其他}\end{cases}$$

2.22 $\boldsymbol{X}$ 是 n 维连续型随机矢量，$\boldsymbol{Y}=\boldsymbol{AX}$ 是 $\boldsymbol{X}$ 的线性变换，并且 $\boldsymbol{A}$ 是一个 $n\times n$ 的非奇异矩阵。证明：$h(\boldsymbol{Y})=\log|\det(\boldsymbol{A})|+h(\boldsymbol{X})$。

2.23 设以 8000 样值/秒的速率抽样一语音信号，并以 $M=2^8=256$ 级对抽样均匀量化，设抽样值取各量化值的概率相等，且抽样间相互统计独立。求：

（1）求每抽样的信息熵。

（2）求信源的信息输出率。

第 3 章　信道及其信道容量

信道是指信息传输的通道，包括空间传输和时间传输。我们在实际通信中所利用的各种物理通道是空间传输信道最典型的例子，如电缆、光纤、电波的传输空间、载波线路等；时间传输是指将信息保存，在以后读取，如磁带、光盘等在时间上将信息进行传输的信道。有时我们将为了某种目的而使信息不得不经过的通道视为信道，如一个分类器的输入到输出就可以视为一个信道。这里最关键的是信道有一个输入以及一个与输入有关的输出。信道本身的物理结构可能是千差万别的，最简单的如一个放大器的输入到输出，而复杂的如一条国际通信线路，其中可能包括终端设备、电缆、微波等。信息论研究的信道，其输入点和输出点在一个实际物理通道中所处位置的选择完全取决于研究的目的。例如，通信中我们可以把发送天线到接收天线之间的通道视为信道，也可以把从话机到话机之间的通道视为信道。

关于信道的主要问题如下：① 信道的建模（信道的统计特性的描述）；② 信道容量的计算；③ 在有噪信道中能不能实现可靠传输？怎样实现可靠传输？本章将回答前两个问题，第 5 章介绍第三个问题。下面按信道的分类介绍它们的数学模型及信道容量的计算。

3.1　信道的分类

在通信中，信道按其物理组成常被分成微波信道、光纤信道、电缆信道等，这种分类是因为信号在这些信道中传输时遵循不同的物理规律，而通信技术必须研究这些规律以获得信号在这些信道中的传输特性。信息论不研究怎样获得这些传输特性，而假定传输特性是已知的，并在此基础上研究信息的传输问题。

信息论不研究信号在信道中传输的物理过程，它假定信道的传输特性是已知的，这样信道就可以用图 3.1 所示的抽象数学模型来描述。由于信道输入随机变量 X 和输出随机变量 Y 往往不是确定关系，在信息论中，信道用在输入已知的情况下，输出的条件概率分布 $P(Y|X)$ 来表示，加上输入随机变量 X 和输出随机变量 Y，通常表示成 $\{X, P(Y|X), Y\}$

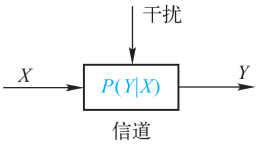


图 3.1　信道模型

根据实际应用的需要，信道有如下几种分类方法。

（1）按其输入/输出信号在幅度和时间上的取值是离散的或连续的来划分

这种分类如表 3.1 所示。

表 3.1　按输入/输出信号在幅度和时间上是离散的或连续的划分

幅　度	时　间	信 道 名 称
离散	离散	离散信道（数字信道）
连续	离散	连续信道
连续	连续	模拟信道（波形信道）
离散	连续	（理论和实用价值均很小）

（2）按其输入/输出信号之间关系的记忆特性来划分

分为有记忆信道和无记忆信道。如果信道的输出只与信道该时刻的输入有关，而与其他时刻的输入无关，则称此信道是无记忆的，反之称为有记忆的。

实际信道一般都是有记忆的，信道中的记忆现象来源于物理信道中的惯性，如电缆信道中的电感电容、无线信道中的电波传播的衰落现象等。有记忆信道的分析比较复杂，有用的研究成果很少，因此主要研究无记忆信道。

（3）按输入/输出信号之间的关系是否确定来划分

分为有噪声信道和无噪声信道。一般来说，因为信道中总是存在某种程度的噪声，所以信道输入/输出之间的关系是一种统计依存的关系。但是，如果噪声与信号相比很小，则可以近似为无噪声信道。有噪声信道是信息论研究的主要对象。信道输入、输出及信道输入/输出信号之间的统计关系的描述，就构成了有噪声信道的数学模型。

（4）根据信道输入和输出的个数来划分

两端信道（单用户信道）：只有一个输入端和一个输出端的单向通信的信道。

多端信道（多用户信道）：双向通信或三个或更多个用户之间相互通信的情况。

本课程主要研究两端信道的情况。

（5）根据信道的统计特性是否随时间变化来划分

恒参信道（平稳信道）：信道的统计特性不随时间变化。卫星通信信道在某种意义下可以近似为恒参信道。

随参信道（非平稳信道）：信道的统计特性随时间变化。如短波通信中，其信道可视为随参信道。

本书主要研究恒参信道的情况。

3.2　离散单符号信道

3.2.1　离散单符号信道的数学模型

信道的输入、输出都取值于离散符号集，都用一个随机变量来表示的信道，就是离散单符号信道。

设离散单符号信道的输入随机变量为 X，其所有可能的取值为 $x_i(i=1,2,\cdots,r)$，输出随机变量为 Y，其所有可能的取值为 $y_j(j=1,2,\cdots,s)$，由于信道中存在干扰，因此输入符号在传输中将会产生错误，这种信道干扰对传输的影响可用传递概率 $p(y_j \mid x_i)$ 来描述：

$$p(y_j \mid x_i) = P(Y = y_j \mid X = x_i) \qquad i = 1,2,\cdots,r;j = 1,2,\cdots,s$$

信道传递概率实际上是一个传递概率矩阵，称为信道矩阵，记为

$$\boldsymbol{P} = \begin{array}{c} \\ x_1 \\ x_2 \\ \vdots \\ x_r \end{array} \begin{array}{c} \begin{array}{cccc} y_1 & y_2 & \cdots & y_s \end{array} \\ \begin{bmatrix} p(y_1 \mid x_1) & p(y_2 \mid x_1) & \cdots & p(y_s \mid x_1) \\ p(y_1 \mid x_2) & p(y_2 \mid x_2) & \cdots & p(y_s \mid x_2) \\ \vdots & \vdots & \ddots & \vdots \\ p(y_1 \mid x_r) & p(y_2 \mid x_r) & \cdots & p(y_s \mid x_r) \end{bmatrix} \end{array}$$

为了方便表述，常常将信道矩阵记为

$$\boldsymbol{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1s} \\ p_{21} & p_{22} & \cdots & p_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ p_{r1} & p_{r2} & \cdots & p_{rs} \end{bmatrix}$$

并且传递概率满足

$$p_{ij} \geqslant 0$$

$$\sum_{j=1}^{s} p_{ij} = 1 \qquad i = 1,2,\cdots,r$$

即信道矩阵中每个元素均为非负，每行元素之和为 1。

最常见的信道是二元对称信道（Binary Symmetric Channel，BSC），如图 3.2 所示，输入符号集和输出符号集分别为 $X = \{0,1\}$ 和 $Y = \{0,1\}$。其信道传递概率为

$$p(y_1 \mid x_1) = p(0 \mid 0) = 1 - p = \bar{p}$$

$$p(y_2 \mid x_1) = p(1 \mid 0) = p$$

$$p(y_2 \mid x_2) = p(1 \mid 1) = 1 - p = \bar{p}$$

$$p(y_1 \mid x_2) = p(0 \mid 1) = p$$

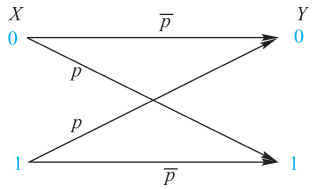


图 3.2　二元对称信道

式中，$\bar{p}$表示单个符号无错误传输的概率，p 表示单个符号传输发生错误的概率。所以，二元对称信道的信道矩阵为

$$\boldsymbol{P} = \begin{bmatrix} \bar{p} & p \\ p & \bar{p} \end{bmatrix}$$

满足

$$\sum_{j=1}^{2} p(y_j \mid x_1) = \sum_{j=1}^{2} p(y_j \mid x_2) = 1$$

下面推导一般离散单符号信道的一些概率关系。设信道输入随机变量的概率空间为

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{bmatrix} x_1 & \cdots & x_2 & \cdots & x_r \\ p(x_1) & \cdots & p(x_2) & \cdots & p(x_r) \end{bmatrix}$$

并且 $\sum_{i=1}^{r} p(x_i) = 1$，$0 \leqslant p(x_i) \leqslant 1 (i = 1,2,\cdots,r)$。

再设信道输出随机变量的概率空间为

$$\begin{bmatrix} Y \\ P(Y) \end{bmatrix} = \begin{bmatrix} y_1 & \cdots & y_2 & \cdots & y_s \\ p(y_1) & \cdots & p(y_2) & \cdots & p(y_s) \end{bmatrix}$$

并且 $\sum_{j=1}^{s} p(y_j) = 1$，$0 \leqslant p(y_j) \leqslant 1 (j = 1, 2, \cdots, s)$。

给定信道矩阵为

$$\boldsymbol{P} = \begin{bmatrix} p(y_1 \mid x_1) & p(y_2 \mid x_1) & \cdots & p(y_s \mid x_1) \\ p(y_1 \mid x_2) & p(y_2 \mid x_2) & \cdots & p(y_s \mid x_2) \\ \vdots & \vdots & \ddots & \vdots \\ p(y_1 \mid x_r) & p(y_2 \mid x_r) & \cdots & p(y_s \mid x_r) \end{bmatrix}$$

（1）输入/输出随机变量的联合概率分布为 $p(x_i y_j) = P(X = x_i, Y = y_j)$，则

$$p(x_i y_j) = p(x_i) p(y_j \mid x_i) = p(y_j) p(x_i \mid y_j) \tag{3.1}$$

其中，$p(y_j \mid x_i)$是信道传递概率，即输入为 x_i，通过信道传输输出 y_j 的概率，通常称为前向概率。它是由信道噪声引起的，所以通常用它描述信道噪声的特性。而 $p(x_i \mid y_j)$是已知信道输出符号 y_j，输入符号为 x_i 的概率，称为后向概率。有时，$p(x_i)$被称为输入符号的先验概率（在接收到输出符号之前，判断输入符号为 x_i 的概率）。对应地，$p(x_i \mid y_j)$被称为输入符号的后验概率（接收到输出符号 y_j 之后，输入符号为 x_i 的概率）。

（2）由全概率公式，可从先验概率和信道传递概率求输出符号的概率：

$$p(y_j) = \sum_{i=1}^{r} p(x_i) p(y_j \mid x_i) \tag{3.2}$$

写成向量形式为

$$[p(y_1) \quad p(y_2) \quad \cdots \quad p(y_s)] = [p(x_1) \quad p(x_2) \quad \cdots \quad p(x_r)] \cdot \boldsymbol{P}$$

或记为

$$\boldsymbol{P}_Y = \boldsymbol{P}_X \boldsymbol{P}_{Y|X}$$

（3）根据贝叶斯公式，可由先验概率和信道的传递概率求后向概率：

$$\begin{aligned} p(x_i \mid y_j) &= \frac{p(x_i y_j)}{p(y_j)} \\ &= \frac{p(x_i) p(y_j \mid x_i)}{\sum_{i=1}^{r} p(x_i) p(y_j \mid x_i)} \end{aligned} \tag{3.3}$$

$$\sum_{i=1}^{r} p(x_i \mid y_j) = 1$$

其中，$i = 1, 2, \cdots, r$，$j = 1, 2, \cdots, s$。

3.2.2 信道容量的概念

平均互信息 $I(X;Y)$是接收到输出随机变量 Y 后所获得的关于输入随机变量 X 的信息量。信源的不确定性为 $H(X)$，由于干扰的存在，接收端收到 Y 后对信源仍然存在的不确定性为 $H(X \mid Y)$。$H(X \mid Y)$又被称为信道疑义度。信宿所消除的关于信源的不确定性，也就是获得的关于信源的信息为 $I(X;Y)$，它是平均意义上每传送一个符号流经信道的信息量，从

这个意义上来说，平均互信息 $I(X;Y)$ 又称为信道的信息传输率，通常用 R 表示，即

$$R = I(X;Y) = H(X) - H(X \mid Y) \text{ 比特/符号} \tag{3.4}$$

有时人们关心的是信道在单位时间内平均传输的信息量。如果平均传输一个符号为 t 秒，则信道平均每秒钟传输的信息量一般被称为信息传输速率：

$$R_t = \frac{1}{t} I(X;Y) \text{ 比特/秒} \tag{3.5}$$

$I(X;Y)$ 是信源概率分布 $p(x_i)$ 和信道转移概率 $p(y_j \mid x_i)$ 的二元函数，当信道特性 $p(y_j \mid x_i)$ 固定后，$I(X;Y)$ 是 $p(x_i)$ 的一元函数，并且由定理 1－1 可知，对于给定的信道转移概率 $p(y_j \mid x_i)$，$I(X;Y)$ 是输入分布 $p(x_i)$ 的上凸函数。因此，对于固定的信道，总存在一种信源（某种输入概率分布），使信道平均传输一个符号接收端获得的信息量最大。也就是说，对于每个固定信道都有一个最大的信息传输率，这个最大的信息传输率即信道容量，而相应的输入概率分布称为最佳输入分布。

因此，对于某一个固定信道，必然有一个最佳输入分布使 $I(X;Y)$ 得到极大值。信道容量是信道转移概率的函数，是由信道的统计特性决定的，是某个信道的最大信息传输速率。

【定义 3-1】信道容量为平均互信息对于输入概率分布的最大值：

$$C \overset{\text{def}}{=\!=} \max_{p(x)} \{ I(X;Y) \} \tag{3.6}$$

单位依所用的对数的底的不同，可以是比特/符号、奈特/符号等。

若平均传输一个符号需要 t 秒，则信道在单位时间内平均传输的最大信息量

$$C_t \overset{\text{def}}{=\!=} \max_{p(x)} \{ I(X;Y) \} \tag{3.7}$$

信道容量是信道传输信息的最大能力的度量，信道实际传输的信息量必然不大于信道容量。如果待传输的信息量大于信道容量，则在传输过程中将会发生错误。这是信道编码定理即香农第二定理的内容。

下面以二元对称信道为例，说明信道容量与输入概率分布和信道转移概率的关系。

【例 3.1】输入概率分布 $\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \omega & \bar{\omega} \end{bmatrix}$，信道矩阵为 $\boldsymbol{P} = \begin{bmatrix} \bar{p} & p \\ p & \bar{p} \end{bmatrix}$，$p$ 为信道错误传递概率。求二元对称信道的信道容量。

【解】二元对称信道的平均互信息 $I(X;Y) = H(\omega \bar{p} + \bar{\omega} p) - H(p)$。

固定信道时，p 是一个固定常数，$I(X;Y)$ 是输入概率分布 ω 的上凸函数，因此存在一个关于 ω 的极大值。当 $\omega = \bar{\omega} = \frac{1}{2}$ 时，$H(\omega \bar{p} + \bar{\omega} p) = H\left(\frac{1}{2}\right) = 1$，因而二元对称信道的信道容量 $C = 1 - H(p)$ 比特/符号。

由此可见，信道容量 C 仅为信道传递概率 p 的函数，而与信道输入随机变量 X 的概率分布无关。不同的二元对称信道，其传递概率不同，信道容量也不同。

图 3.3 表示不同的二元对称信道，其传递概率 p 不同，信道容量也不同。

当 $p = \frac{1}{2}$ 时，是一种最坏的二元对称信道，这时 $C = 0$，即该信道不能传递任何信息，信息全部损失在信道中了。当 $p = 0$ 或 $p = 1$ 时，$C = 1$，这是最好的情况，信道能够无失真地传送信源信息。

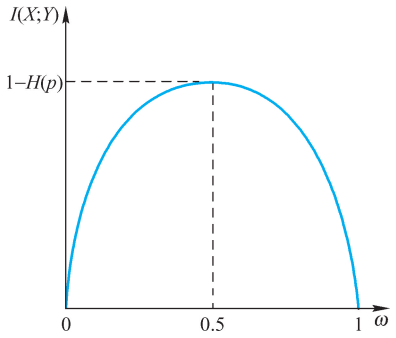


图 3.3　固定二元对称信道的平均互信息

3.2.3　几种特殊信道的信道容量

对于一般信道，求信道容量的计算非常复杂，需要对平均互信息 $I(X;Y)$ 求极大值。下面先讨论某些特殊类型信道的信道容量，然后讨论一般离散信道的信道容量的计算。

① 具有扩展性能的无损信道（如图 3.4 所示），一个输入对应多个输出。

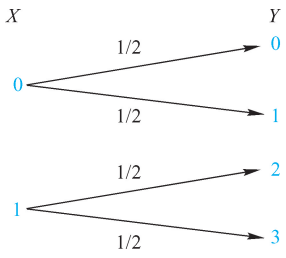


图 3.4　无损信道

例如，信道矩阵

$$\boldsymbol{P}=\begin{bmatrix}\frac{1}{2} & \frac{1}{2} & 0 & 0\\ 0 & 0 & \frac{1}{2} & \frac{1}{2}\end{bmatrix}$$

无损信道信道矩阵中的每列只有一个非零元素，接收到信道输出符号后对输入符号将不存在不确定性，即信道疑义度 $H(X\mid Y)=0$。同时，$H(X\mid Y)$ 表示信源符号通过有噪信道传输后损失的信息量，因为如果没有信息损失，信源包含的信息量将全部到达接收端，接收端将对信源不再有不确定性，所以 $H(X\mid Y)$ 又称为损失熵。对于无损信道，有

$$I(X;Y)=H(X)-H(X\mid Y)=H(X) \tag{3.8}$$

其信道容量为

$$C=\max_{p(x)}\{I(X;Y)\}=\max_{p(x)}H(X)=\log r \tag{3.9}$$

当信道输入等概分布时，信道达到信道容量。由于噪声熵 $H(Y\mid X)>0$，所以

$$I(X;Y)=H(X)<H(Y) \tag{3.10}$$

② 具有归并性能的无噪信道（如图 3.5 所示），一个输出对应多个输入。

例如，信道矩阵

$$\boldsymbol{P}=\begin{bmatrix}1 & 0\\ 1 & 0\\ 0 & 1\\ 0 & 1\end{bmatrix}$$

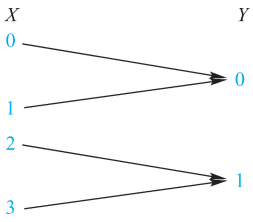


图 3.5　无噪信道

无噪信道的每行只有一个非零元素 1，信道矩阵元素非零即 1。已知信道输入符号，必能确定输出符号，因此 $H(Y\mid X)=0$。

$H(Y|X)$又称为噪声熵，因为是信道的噪声使得$H(Y|X)\neq 0$。

无噪信道的信道容量为

$$C=\max_{p(x)}\{I(X;Y)\}=\max_{p(x)}H(Y)=\log s \tag{3.11}$$

当信道输出等概分布时，信道达到信道容量。

但是输出端接收到某个符号后并不能确定是哪个输入符号，因此信道疑义度$H(X|Y)>0$，于是无噪信道的平均互信息

$$I(X;Y)=H(Y)<H(X) \tag{3.12}$$

③ 具有一一对应关系的无噪无损信道（如图3.6所示），输入、输出之间有确定的一一对应关系，即$y=f(x)$。信道传递概率为

$$p(y_j|x_i)=\begin{cases}1 & y_j=f(x_i)\\ 0 & y_j\neq f(x_i)\end{cases}$$

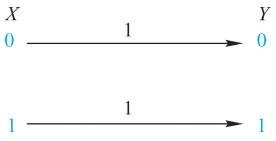


图3.6　无噪无损信道

例如，信道矩阵为

$$\boldsymbol{P}=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$$

无噪无损信道的每行、每列只有一个“1”，已知X后对Y不存在不确定性，收到Y后对X也不存在不确定性，所以噪声熵和损失熵均为0，则

$$I(X;Y)=H(X)=H(Y) \tag{3.13}$$

其信道容量为

$$C=\max_{p(x)}\{I(X;Y)\}=\max_{p(x)}H(Y)=\max_{p(x)}H(X)=\log s=\log r \tag{3.14}$$

当信道输入等概分布时，输出也为等概分布，信道达到信道容量。

对于以上三种信道，求信道容量C的问题已从求$I(X;Y)$的极值问题，转化为求$H(X)$或$H(Y)$的极值问题。信道容量C只决定于信道的输入符号数r或输出符号数s，与信源无关，它表征信道的统计特性。

3.2.4　离散对称信道的信道容量

离散信道中有一类特殊的信道，其特点是信道矩阵具有行对称性，利用这个对称性，我们可以简化信道容量的计算。

【定义3-2】如果信道矩阵中每行都是第一行元素的不同排列，则称此类信道为行对称信道。

例如，

$$\boldsymbol{P}=\begin{bmatrix}\frac{1}{3} & \frac{1}{3} & \frac{1}{6} & \frac{1}{6}\\ \frac{1}{6} & \frac{1}{3} & \frac{1}{6} & \frac{1}{3}\end{bmatrix}$$

和

$$\boldsymbol{P}=\begin{bmatrix}\frac{1}{3} & \frac{1}{3} & \frac{1}{6} & \frac{1}{6}\\ \frac{1}{6} & \frac{1}{6} & \frac{1}{3} & \frac{1}{3}\end{bmatrix}$$

都是行对称信道。

【定义 3-3】若信道矩阵中不但每行都是第一行元素的不同排列，而且每列都是第一列元素的不同排列，这类信道称为对称信道。

例如：

$$\boldsymbol{P}=\begin{bmatrix}\frac{1}{3} & \frac{1}{3} & \frac{1}{6} & \frac{1}{6}\\ \frac{1}{6} & \frac{1}{6} & \frac{1}{3} & \frac{1}{3}\end{bmatrix}$$

$$\boldsymbol{P}=\begin{bmatrix}\frac{1}{2} & \frac{1}{3} & \frac{1}{6}\\ \frac{1}{6} & \frac{1}{2} & \frac{1}{3}\\ \frac{1}{3} & \frac{1}{6} & \frac{1}{2}\end{bmatrix}$$

都是对称信道。

【定义 3-4】若信道矩阵中每行都是第一行元素的不同排列，每列并不都是第一列元素的不同排列，但可以按照信道矩阵的列将信道矩阵划分成若干对称的子矩阵，则称这类信道为离散准对称信道。

例如，信道矩阵

$$\boldsymbol{P}=\begin{bmatrix}0.8 & 0.1 & 0.1\\ 0.1 & 0.1 & 0.8\end{bmatrix}$$

可以划分成两个对称的子矩阵

$$\boldsymbol{P}_1=\begin{bmatrix}0.8 & 0.1\\ 0.1 & 0.8\end{bmatrix}\text{和}\ \boldsymbol{P}_2=\begin{bmatrix}0.1\\ 0.1\end{bmatrix}$$

因此它是准对称信道。

【定义 3-5】若对称信道中输入符号和输出符号的个数相同，且信道中总的错误概率为 p，对称地平均分配给 $r-1$ 个输出符号，r 为输入输出符号的个数，即信道矩阵为

$$\boldsymbol{P}=\begin{bmatrix}\bar{p} & \frac{p}{r-1} & \frac{p}{r-1} & \cdots & \frac{p}{r-1}\\ \frac{p}{r-1} & \bar{p} & \frac{p}{r-1} & \cdots & \frac{p}{r-1}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \frac{p}{r-1} & \frac{p}{r-1} & \frac{p}{r-1} & \cdots & \bar{p}\end{bmatrix}$$

则称此信道为强对称信道或均匀信道。

二元对称信道就是 $r=2$ 的均匀信道。一般信道的信道矩阵中各行之和为 1，但各列之和不一定等于 1，而均匀信道中各列之和亦等于 1。

【定理 3-1】对于对称信道，当输入分布为等概分布时，输出分布必能达到等概分布。

【证明】

当输入为等概分布时，$p(x_i)=\frac{1}{r}(i=1,2,\cdots,r)$，输出

$$p(y_j)=\sum_i p(x_i)p(y_j\mid x_i)=\frac{1}{r}\sum_i p(y_j\mid x_i)=\frac{1}{r}H_j \tag{3.15}$$

式中，$H_j=\sum_{i=1}^{r}p(y_j\mid x_i)$ 表示信道矩阵 $\boldsymbol{P}$ 中第 j 列元素之和。由信道的对称性可知，H_j是一个与 j 无关的常数，每列元素之和均为 H_j。由于信道矩阵中每行的元素之和为 1，所以 r 行元素之和为 r，并且 r 行元素之和必等于 s 列元素之和，即 $sH_j=r$，$H_j=\frac{r}{s}$，因此

$$p(y_j)=\frac{1}{r}H_j=\frac{1}{s}\qquad j=1,2,\cdots,s \tag{3.16}$$

即当信道输入为等概分布 $p(x_i)=\frac{1}{r}$ $(i=1,2,\cdots,r)$时，输出 $p(y_j)=\frac{1}{s}$ $(j=1,2,\cdots,s)$亦为等概分布。

证毕。

【定理 3-2】若一个离散对称信道具有 r 个输入符号，s 个输出符号，则当输入为等概分布时达到信道容量，且

$$C=\log s-H(p'_1,p'_2,\cdots,p'_s) \tag{3.17}$$

其中，$p'_1,p'_2,\cdots,p'_s$ 为信道矩阵中的任一行。

【证明】

平均互信息为

$$I(X;Y)=H(Y)-H(Y\mid X) \tag{3.18}$$

其中，噪声熵为

$$\begin{aligned}H(Y\mid X)&=\sum_i\sum_j p(x_iy_j)\log\frac{1}{p(y_j\mid x_i)}\\&=\sum_i p(x_i)\sum_j p(y_j\mid x_i)\log\frac{1}{p(y_j\mid x_i)}\\&=\sum_i p(x_i)H(Y\mid x_i)\end{aligned}$$

由于信道的对称性，$H(Y\mid x_i)$与 x_i 无关，且 $H(Y\mid x_i)=H(p'_1,p'_2,\cdots,p'_s)$，所以

$$I(X;Y)=H(Y)-H(p'_1,p'_2\cdots,p'_s)$$

根据信道容量的定义，可得

$$\begin{aligned}C&=\max_{p(x)}\{I(X;Y)\}\\&=\max_{p(x)}\{H(Y)-H(p'_1,p'_2,\cdots,p'_s)\}\\&=\max_{p(x)}H(Y)-H(p'_1,p'_2,\cdots,p'_s)\end{aligned}$$

当输出 Y 为等概分布时，$H(Y)$达到最大 $\log s$。所以，当信源 X 的概率分布使输出 Y 等概分布时，信道达到信道容量，且$C=\log s-H(p'_1,p'_2,\cdots,p'_s)$，即信道容量只与输出符号个数和信道矩阵中的任一行元素 $p'_1,p'_2,\cdots,p'_s$ 有关。

【推论 3-1】均匀信道的信道容量为

$$C=\log r-p\log(r-1)-H(p) \tag{3.19}$$

【证明】

均匀信道中输入、输出符号数相等，即 $r=s$，所以

$$
\begin{aligned}
C &= \log r - H(p'_1, p'_2, \cdots, p'_s) \\
&= \log r - H\left(\bar{p}, \frac{p}{r-1}, \cdots, \frac{p}{r-1}\right) \\
&= \log r + \bar{p}\log\bar{p} + \frac{p}{r-1}\log\frac{p}{r-1} + \cdots + \frac{p}{r-1}\log\frac{p}{r-1} \\
&= \log r + \bar{p}\log\bar{p} + p\log\frac{p}{r-1} \\
&= \log r - p\log(r-1) + \bar{p}\log\bar{p} + p\log p \\
&= \log r - p\log(r-1) - H(p)
\end{aligned}
$$

其中，p 是总的错误传递概率，$\bar{p}$是正确传递概率。当输入为等概分布时，输出为等概分布，信道达到信道容量。$r=2$ 时的均匀信道常称为二元对称信道，这时 $C = 1 - H(p)$。

对于一般的离散行对称信道，信道容量 C 仍然可以写成

$$
C = \max_{p(x)}\{H(Y)\} - H(p'_1, p'_2, \cdots, p'_s) \tag{3.20}
$$

但不一定存在一种输入分布能使输出达到等概分布，此时的信道容量

$$
C \leqslant \log s - H(p'_1, p'_2, \cdots, p'_s) \tag{3.21}
$$

而离散对称信道的信道矩阵中的每列都是由同一组元素的不同排列组成的，所以保证了当输入符号 X 为等概分布时，输出符号 Y 一定也是等概分布，输出随机变量熵可以达到 $\log s$。

对于离散准对称信道，由于不一定存在一种输入分布使输出等概，从而

$$
C \leqslant \log s - H(p'_1, p'_2, \cdots, p'_s)
$$

其是可以证明当输入为等概分布时，可以达到信道容量

$$
C = \log r - \sum_{k=1}^{n} N_k \log M_k - H(p'_1, p'_2, \cdots, p'_s) \tag{3.22}
$$

其中，N_k 是第 k 个子矩阵中的行元素之和，M_k 是第 k 个子矩阵中的列元素之和。（证明留给读者。）

【例 3.2】设某离散对称信道的信道矩阵为

$$
\boldsymbol{P} = \begin{bmatrix}
\frac{1}{2} & \frac{1}{3} & \frac{1}{6} \\
\frac{1}{6} & \frac{1}{2} & \frac{1}{3} \\
\frac{1}{3} & \frac{1}{6} & \frac{1}{2}
\end{bmatrix}
$$

求信道容量。

【解】

这是一个对称信道，有

$$
\begin{aligned}
C &= \log s - H(p'_1, p'_2, \cdots, p'_s) \\
&= \log 3 - H\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right) \\
&= \log 3 + \frac{1}{2}\log\frac{1}{2} + \frac{1}{3}\log\frac{1}{3} + \frac{1}{6}\log\frac{1}{6} \\
&= 0.126 \text{ 比特/符号}
\end{aligned}
$$

在这个对称信道中，每个符号平均能够传输的最大信息量为 0.126 比特。只有当信道输入符号是等概分布时可以达到这个最大值。

【例 3.3】求二元对称删除信道的信道容量。

$$\boldsymbol{P}=\begin{bmatrix}1-p-q & q & p\\ p & q & 1-p-q\end{bmatrix}$$

【解】

这是一个准对称信道，则

$$N_1=1-q,\ M_1=1-q,\ N_2=q,\ M_2=2q$$

$$\begin{aligned}C&=\log r-\sum N_k\log M_k-H(p_1,p_2,\cdots,p_s)\\&=\log 2-(1-q)\log(1-q)-q\log(2q)-H(1-p-q,q,p)\end{aligned}$$

3.2.5 一般离散信道的信道容量

信道容量定义为在信道固定的条件下，求平均互信息对所有可能的输入分布的极大值。前面已经导出，平均互信息 $I(X;Y)$ 是输入概率分布 $p(x)$ 的上凸函数，因此极大值必定存在。

在信道固定的条件下，平均互信息 $I(X;Y)$ 是 r 个变量 $p(x_i)(i=1,2,\cdots,r)$ 的多元函数，且满足约束条件

$$\sum_{i=1}^{r}p(x_i)=1$$

故可用拉格朗日乘子法来求这个条件极值，即在

$$\begin{cases}p(x_i)\geqslant 0\\ \sum\limits_i p(x_i)=1\end{cases}\qquad i=1,2,\cdots,r$$

条件下求 $I(X;Y)$ 的极值。因为 $I(X;Y)$ 是关于 $p(x_i)$ 的上凸函数，所以得到的极值是极大值。

设辅助函数

$$F=I(X;Y)-\lambda\sum_i p(x_i)\tag{3.23}$$

当 $\dfrac{\partial F}{\partial p(x_i)}=0$ 时，求得的 $I(X;Y)$ 的极值即为信道容量。

对式（3.23）进行整理，有

$$\begin{aligned}F&=H(y)-H(y\mid x)-\lambda\sum_i p(x_i)\\&=\sum_i p(x_i)\sum_j p(y_j\mid x_i)\log p(y_j\mid x_i)-\sum_j p(y_j)\log p(y_j)-\lambda\sum_i p(x_i)\end{aligned}\tag{3.24}$$

因为

$$p(y_j)=\sum_i p(x_i)p(y_j\mid x_i)\tag{3.25}$$

所以

$$\frac{\partial p(y_j)}{\partial p(x_i)}=p(y_j\mid x_i)\tag{3.26}$$

又因为

$$\log p(y_j)=\frac{\ln p(y_j)}{\ln 2} \tag{3.27}$$

所以

$$\begin{aligned}\frac{\partial \log p(y_j)}{\partial p(x_i)}&=\frac{\partial \ln p(y_j)}{\partial p(x_i)}\cdot\frac{1}{\ln 2}\\&=\frac{1}{p(y_j)}\frac{\partial p(y_j)}{\partial p(x_i)}\log e\\&=\frac{1}{p(y_j)}p(y_j\mid x_i)\log e\\&=\frac{p(y_j\mid x_i)}{p(y_j)}\log e\end{aligned} \tag{3.28}$$

因此有

$$\begin{aligned}\frac{\partial F}{\partial p(x_i)}&=\sum_j p(y_j\mid x_i)\log p(y_j\mid x_i)-\sum_j p(y_j\mid x_i)\log p(y_j)-\sum_j p(y_j)\frac{p(y_j\mid x_i)}{p(y_j)}\log e-\lambda\\&=\sum_j p(y_j\mid x_i)\log\frac{p(y_j\mid x_i)}{p(y_j)}-\sum_j p(y_j)\frac{p(y_j\mid x_i)}{p(y_j)}\log e-\lambda\\&=\sum_j p(y_j\mid x_i)\log\frac{p(y_j\mid x_i)}{p(y_j)}-\log e-\lambda\end{aligned} \tag{3.29}$$

令$\frac{\partial F}{\partial p(x_i)}=0$，则

$$\sum_j p(y_j\mid x_i)\log\frac{p(y_j\mid x_i)}{p(y_j)}-\log e-\lambda=0 \tag{3.30}$$

即

$$\sum_j p(y_j\mid x_i)\log\frac{p(y_j\mid x_i)}{p(y_j)}=\log e+\lambda\qquad i=1,2,\cdots,r \tag{3.31}$$

将式（3.31）两边同乘以$p(x_i)$并对i求和，有

$$\sum_i p(x_i)\sum_j p(y_j\mid x_i)\log\frac{p(y_j\mid x_i)}{p(y_j)}=\sum_i p(x_i)(\log e+\lambda) \tag{3.32}$$

式（3.32）左边即为平均互信息的极大值C，所以得到

$$C=\log e+\lambda \tag{3.33}$$

这样得到的信道容量有一个参数λ。在某些情况下，可以消去λ，得到信道容量值。

① 当输入概率分布只有一个变量时，如$r=2$，可以设输入概率分布为α和$1-\alpha$，因此输入概率分布只有一个变量，这时可以直接对$I(X;Y)$求导求出α，从而得出$I(X;Y)$的极大值C。对于$r\neq 2$的情况，可以通过已知条件消去一些变量，使得最后的输入概率分布只有一个变量。

【例3.4】已知信道的转移矩阵为

$$\boldsymbol{P}=\begin{bmatrix}0.5 & 0.3 & 0.2\\0.3 & 0.5 & 0.2\end{bmatrix}$$

求信道容量。

【解】设输入概率分布$p(x_1)=\alpha, p(x_2)=1-\alpha$，则输出$y_1$、$y_2$、$y_3$的概率分布为

$$\begin{aligned}\boldsymbol{P}_Y &= \boldsymbol{P}_X\boldsymbol{P}_{Y|X}\\ &= [\alpha \quad 1-\alpha]\begin{bmatrix}0.5 & 0.3 & 0.2\\ 0.3 & 0.5 & 0.2\end{bmatrix}\\ &= [0.3+0.2\alpha \quad 0.5-0.2\alpha \quad 0.2]\end{aligned}$$

其中，$p(y_3)$固定，与x_i的分布无关。

$$\begin{aligned}I(X;Y) &= H(Y) - H(Y|X)\\ &= -\sum_j p(y_j)\log p(y_j) + \sum_i p(x_i)\sum_j p(y_j|x_i)\log p(y_j|x_i)\\ &= -(0.3+0.2\alpha)\log(0.3+0.2\alpha) - (0.5-0.2\alpha)\log(0.5-0.2\alpha) -\\ &\quad 0.2\log 0.2 + 0.5\log 0.5 + 0.3\log 0.3 + 0.2\log 0.2\end{aligned}$$

由$\dfrac{\partial I(X;Y)}{\partial \alpha}=0$得

$$0.2\log(0.3+0.2\alpha) - 0.2 + 0.2\log(0.5-0.2\alpha) + 0.2 = 0$$

解得$\alpha=\dfrac{1}{2}$，即输入等概分布时$I(X;Y)$达到极大值，且

$$C = \max I(X;Y) = 0.036 \text{ 比特/符号}$$

② 对于信道矩阵为可逆矩阵的信道，我们可以采用解方程组的方法。

在一般信道的信道容量的推导中，我们推出了下式：

$$\sum_j p(y_j|x_i)\log\frac{p(y_j|x_i)}{p(y_j)} = \log e + \lambda = C \qquad i=1,2,\cdots,r \tag{3.34}$$

移项得

$$\begin{aligned}\sum_j p(y_j|x_i)\log p(y_j|x_i) &= \sum_j p(y_j|x_i)\log p(y_j) + C\\ &= \sum_j p(y_j|x_i)[\log p(y_j) + C]\end{aligned} \tag{3.35}$$

令

$$\beta_j = \log p(y_j) + C \tag{3.36}$$

则

$$\sum_j p(y_j|x_i)\log p(y_j|x_i) = \sum_j p(y_j|x_i)\beta_j \tag{3.37}$$

这是含有s个未知数β_j、由r个方程组成的方程组。

当$r=s$，且信道矩阵是可逆矩阵时，该方程组有唯一解。这时就可以求出β_j，然后根据$p(y_j)=2^{\beta_j - C}$和$\sum_j p(y_j)=1$，求出信道容量：

$$\sum_j 2^{\beta_j - C} = 1 \tag{3.38}$$

$$C = \log\sum_j 2^{\beta_j} \tag{3.39}$$

由β_j和C可以求得输出概率分布$p(y_j)$：

$$p(y_j) = 2^{\beta_j - C} \tag{3.40}$$

再根据

$$p(y_j) = \sum_i p(x_i)p(y_j|x_i) \tag{3.41}$$

列方程组求$p(x_i)$。

计算步骤总结如下：

① 由式（3.37）列方程组求出 β_j。

② 由式（3.39）求出 C。

③ 由式（3.40）求出 $p(y_j)$。

④ 由式（3.41）列方程组求 $p(x_i)$。

需要强调的是，在第②步求出信道容量后，计算并未结束，还需解出 $p(x_i)$，如果所有的 $p(x_i)\geqslant 0$，则求出的信道容量才是正确的。这是因为用拉格朗日乘子法没有加入 $p(x_i)\geqslant 0$（$i=1,2,\cdots,r$）的约束条件，因此算出的 $p(x_i)$有可能是负值。如果 $p(x_i)$有负值，则此解无效，它表明所求得的极限值出现的区域不满足概率条件，这时最大值必在边界上，即有某些输入符号的概率 $p(x_i)=0$。因此必须设某些输入符号的概率 $p(x_i)=0$，然后重新进行计算。这样的计算比较复杂，一般要通过迭代来实现。

【例 3.5】求如下信道的信道容量。

$$\boldsymbol{P}=\begin{bmatrix}\frac{1}{2} & \frac{1}{4} & 0 & \frac{1}{4}\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{2}\end{bmatrix}$$

【解】

这个信道矩阵中 $r=s$ 且为可逆矩阵（满秩矩阵），所以以下方程组有唯一解：

$$\begin{cases}\frac{1}{2}\beta_1+\frac{1}{4}\beta_2+\frac{1}{4}\beta_4=\frac{1}{2}\log\frac{1}{2}+\frac{1}{4}\log\frac{1}{4}+\frac{1}{4}\log\frac{1}{4}\\ \beta_2=0\\ \beta_3=0\\ \frac{1}{4}\beta_1+\frac{1}{4}\beta_3+\frac{1}{2}\beta_4=\frac{1}{4}\log\frac{1}{4}+\frac{1}{4}\log\frac{1}{4}+\frac{1}{2}\log\frac{1}{2}\end{cases}$$

解方程组得

$$\begin{aligned}&\beta_2=\beta_3=0\\ &\beta_1=\beta_4=-2\\ &C=\log\sum_j 2^{\beta_j}\\ &\quad=\log(2^{-2}+2^0+2^0+2^{-2})\\ &\quad=\log 5-1\end{aligned}$$

再根据 $p(y_j)=2^{\beta_j-C}$，求 $p(y_j)$：

$$p(y_1)=p(y_4)=2^{-2-\log 5+1}=\frac{1}{10}$$

$$p(y_2)=p(y_3)=2^{0-\log 5+1}=\frac{4}{10}$$

最后根据 $p(y_j)=\sum_i p(x_i)p(y_j\mid x_i)$，列方程组求 $p(x_i)$，求出最佳输入分布：

$$p(x_1)=p(x_4)=\frac{4}{30}$$

$$p(x_2)=p(x_3)=\frac{11}{30}$$

上述求得的 $p(x_i)(i=1,2,3,4)$ 都大于 0，故求得的结果是正确的。

3.2.6 信道容量定理

从以上讨论可知，求信道容量的问题实际上是在约束条件下求多元函数极值的问题，通常情况下，计算量是非常大的。下面介绍一般离散信道的平均互信息 $I(X;Y)$ 达到信道容量的充要条件，在某些情况下，这可以帮助我们较快地找到极值点。

【定理 3-3】设有一般离散信道，它有 r 个输入信号，s 个输出信号。当且仅当存在常数 C 使输入分布 $p(x_i)$ 满足

① $I(x_i;Y)=C \qquad p(x_i)\neq 0$

② $I(x_i;Y)\leqslant C \qquad p(x_i)=0$

时，$I(X;Y)$ 达到最大值。其中，

$$I(x_i;Y)=\sum_j p(y_j\mid x_i)\log\frac{p(y_j\mid x_i)}{p(y_j)} \tag{3.42}$$

它表示信道输入 x_i 时，所给出关于输出 Y 的信息量。常数 C 为所求的信道容量。

在一般离散信道的信道容量的推导中，我们已经得出以下关系式：

$$\frac{\partial I(X;Y)}{\partial p(x_i)}=\sum_j p(y_j\mid x_i)\log\frac{p(y_j\mid x_i)}{p(y_j)}-\log e=I(x_i;Y)-\log e \tag{3.43}$$

根据式（3.43）和式（3.33），可以将上述充要条件改写成

① $\dfrac{\partial I(X;Y)}{\partial p(x_i)}=\lambda \qquad p(x_i)\neq 0$

② $\dfrac{\partial I(X;Y)}{\partial p(x_i)}\leqslant\lambda \qquad p(x_i)=0$

我们将利用信道容量定理引理（见本章扩展阅读）来证明这个改写后的充要条件。

【证明】

（1）充分性，也就是证明如果输入分布 $\boldsymbol{P}=(p_1,p_2,\cdots,p_r)$ 满足

① $\dfrac{\partial I(X;Y)}{\partial p(x_i)}=\lambda \qquad p(x_i)\neq 0$

② $\dfrac{\partial I(X;Y)}{\partial p(x_i)}\leqslant\lambda \qquad p(x_i)=0$

那么 $\boldsymbol{P}$ 一定使平均互信息 $I(X;Y)$ 达到极大值 $I(\boldsymbol{P})$，即对于任何其他输入分布 $\boldsymbol{Q}=(q_1,q_2,\cdots,q_r)$，必然有

$$I(\boldsymbol{Q})\leqslant I(\boldsymbol{P}) \tag{3.44}$$

因为平均互信息 $I(X;Y)$ 是输入分布的上凸函数，所以

$$\theta I(\boldsymbol{Q})+\bar{\theta}I(\boldsymbol{P})\leqslant I(\theta\boldsymbol{Q}+\bar{\theta}\boldsymbol{P}) \tag{3.45}$$

其中，$\theta+\bar{\theta}=1$，$0<\theta<1$。移项得

$$I(\boldsymbol{Q})-I(\boldsymbol{P})\leqslant\frac{I(\theta\boldsymbol{Q}+\bar{\theta}\boldsymbol{P})-I(\boldsymbol{P})}{\theta} \tag{3.46}$$

上式对一切 $0<\theta<1$ 均成立。取 $\theta\to 0$，根据引理，可得

$$I(\boldsymbol{Q})-I(\boldsymbol{P})\leqslant\sum_{i=1}^{r}(q_i-p_i)\frac{\partial I(\boldsymbol{P})}{\partial p_i} \tag{3.47}$$

式中，$p_i=p(x_i)$，$q_i=q(x_i)$。

根据假设，输入分布 $\boldsymbol{P}$ 满足

① $\dfrac{\partial I(\boldsymbol{P})}{\partial p_i}=\lambda \qquad p_i\neq 0$

② $\dfrac{\partial I(\boldsymbol{P})}{\partial p_i}\leqslant\lambda \qquad p_i=0$

所以

$$\begin{aligned} I(\boldsymbol{Q})-I(\boldsymbol{P})&\leqslant\lambda\sum_{i=1}^{r}(q_i-p_i)\\ &=\lambda\left(\sum_{i=1}^{r}q_i-\sum_{i=1}^{r}p_i\right)=0 \end{aligned} \tag{3.48}$$

即

$$I(\boldsymbol{Q})\leqslant I(\boldsymbol{P})$$

充分性得证。

（2）必要性，就是证明如果输入分布 $\boldsymbol{P}$ 使平均互信息 $I(X;Y)$ 达到极大值 $I(\boldsymbol{P})$，则输入分布 $\boldsymbol{P}$ 必然满足

① $\dfrac{\partial I(\boldsymbol{P})}{\partial p_i}=\lambda \qquad p_i\neq 0$

② $\dfrac{\partial I(\boldsymbol{P})}{\partial p_i}\leqslant\lambda \qquad p_i=0$

如果输入分布 $\boldsymbol{P}$ 使平均互信息 $I(X;Y)$ 达到极大值，取任一其他输入分布 $\boldsymbol{Q}=(q_1,q_2,\cdots,q_r)$，必然有

$$I(\theta\boldsymbol{Q}+\bar{\theta}\boldsymbol{P})-I(\boldsymbol{P})\leqslant 0 \tag{3.49}$$

式(3.49)两边同除以 θ，并取 $\theta\to 0$ 时的极限

$$\lim_{\theta\to 0}\frac{1}{\theta}\{I[\theta\boldsymbol{Q}+(1-\theta)\boldsymbol{P}]-I(\boldsymbol{P})\}\leqslant 0 \tag{3.50}$$

根据引理，可得

$$\sum_{i=1}^{r}(q_i-p_i)\frac{\partial I(\boldsymbol{P})}{\partial p_i}\leqslant 0 \tag{3.51}$$

对于输入分布 $\boldsymbol{P}$，因为概率分布的完备性 $\sum_{i=1}^{r}p_i=1$，所以其中至少有一个分量不为零，令 $p_l\neq 0$，再选择另一个输入分布 $\boldsymbol{Q}=(q_1,q_2,\cdots,q_r)$，并且满足

$$\begin{cases} q_l=p_l-\varepsilon \\ q_j=p_j+\varepsilon & \text{（保证输入分布 } \boldsymbol{Q} \text{ 的完备性）} \\ q_i=p_i & \text{（其他分量均相同）} \end{cases} \tag{3.52}$$

其中，ε 为任意数。代入式(3.51)，得到

$$-\varepsilon\frac{\partial I(\boldsymbol{P})}{\partial p_l}+\varepsilon\frac{\partial I(\boldsymbol{P})}{\partial p_j}\leqslant 0 \tag{3.53}$$

令$\frac{\partial I(\boldsymbol{P})}{\partial p_l}=\lambda$，则$\varepsilon\frac{\partial I(\boldsymbol{P})}{\partial p_j}\leqslant\lambda\varepsilon$。因为概率的非负性，所以$p_l-\varepsilon\geqslant 0$，$p_j+\varepsilon\geqslant 0$，$\varepsilon$必满足$-p_j\leqslant\varepsilon\leqslant p_l$。

当$p_j=0$时，$0\leqslant\varepsilon\leqslant p_l$，$\varepsilon$为正数，所以$\frac{\partial I(\boldsymbol{P})}{\partial p_j}\leqslant\lambda$；当$p_j\neq 0$时，则$-p_j\leqslant\varepsilon\leqslant p_l$，$\varepsilon$可为正数，也可为负数。

如果ε取正数，则$\frac{\partial I(\boldsymbol{P})}{\partial p_j}\leqslant\lambda$；如果$\varepsilon$取负数，则$\frac{\partial I(\boldsymbol{P})}{\partial p_j}\geqslant\lambda$。所以，当$p_j\neq 0$时，必然有$\frac{\partial I(\boldsymbol{P})}{\partial p_j}=\lambda$，因此输入分布$\boldsymbol{P}$必满足充要条件。

必要性得证。

证毕。

$I(x_i;Y)$表示信道输入x_i时，所给出关于输出Y的信息量。一般，x_i不同，$I(x_i;Y)$值也不同。信道容量定理告诉我们，平均互信息$I(X;Y)$取到极大值也就是信道容量时，对于任意x_i，只要它出现的概率大于0，$I(x_i;Y)$都相等。

【例3.6】证明当输入为等概分布时，准对称离散无记忆信道达到信道容量。

【证明】根据信道容量定理，需要证明输入为等概分布$p(x_i)=\frac{1}{r}$时，$I(x_i;Y)$为一个与x_i无关的常数。

$$\begin{aligned}I(x_i;Y)&=\sum_{j=1}^{s}p(y_j\mid x_i)\log\frac{p(y_j\mid x_i)}{p(y_j)}\\&=\sum_{j=1}^{s}p(y_j\mid x_i)\log\frac{p(y_j\mid x_i)}{\sum_{k=1}^{r}p(x_k)p(y_j\mid x_k)}\\&=\sum_{j=1}^{s}p(y_j\mid x_i)\log\frac{p(y_j\mid x_i)}{\frac{1}{r}\sum_{k=1}^{r}p(y_j\mid x_k)}\end{aligned}$$

准对称信道的信道矩阵可按列分为一些对称的子阵$\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_l,\cdots,\boldsymbol{P}_n$。在同一子阵中，每列都是第一列的同一组元素的排列，所以在同一子阵$\boldsymbol{P}_l$中，$p(y_j)=\frac{1}{r}\sum_{k=1}^{r}p(y_j\mid x_k)$，$y_j\in Y_l$都相等。而同一子阵中的每行都是其他行的同一组元素的排列，所以同一子阵$\boldsymbol{P}_l$中，对于任意x_i，$\sum_{y_j\in Y_l}p(y_j\mid x_i)\log\frac{p(y_j\mid x_i)}{\frac{1}{r}\sum_{k=1}^{r}p(y_j\mid x_i)}$都相等。于是，对于任意$x_i$，则必然有

$$I(x_i;Y)=\sum_{l}\sum_{y_j\in Y_l}p(y_j\mid x_i)\log\frac{p(y_j\mid x_i)}{\frac{1}{r}\sum_{k=1}^{r}p(y_j\mid x_i)}$$

所以，对于任意x_i，$I(x_i;Y)$是一个与x_i无关的常数，根据信道容量定理，这时信道达到信道容量，即当输入为等概分布时，准对称离散无记忆信道达到信道容量。

证毕。

信道容量定理只给出了达到信道容量时，最佳输入概率分布应满足的条件，并没有给出

最佳输入概率分布值，也没有给出信道容量的数值。另外，定理本身也隐含着达到信道容量的最佳分布不一定是唯一的，只要输入概率分布满足充要条件式，并使 $I(\boldsymbol{P})$ 最大，就是信道的最佳输入分布。在一些特殊情况下，我们常利用这一定理寻求输入分布和信道容量值。

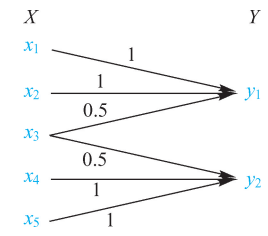


图 3.7　例 3.7 的离散信道

【例 3.7】设离散信道如图 3.7 所示，输入符号集 $\{x_1,x_2,x_3,x_4,x_5\}$，输出符号集 $\{y_1,y_2\}$，求 C。

【解】

信道矩阵 $\boldsymbol{P}=\begin{bmatrix}1 & 0\\ 1 & 0\\ \frac{1}{2} & \frac{1}{2}\\ 0 & 1\\ 0 & 1\end{bmatrix}$，这个信道不是对称信道。由于 x_3 传递到 y_1、y_2 是等概的，如果令 $p(x_3)=0$，则会减少收到 Y 以后对输入 X 的不确定性，这时 x_1、x_2 与 x_4、x_5 分别转移到 y_1、y_2。如果令 $p(x_2)=p(x_4)=0$，则信道就变成了一一对应的信道，接收到 Y 后对输入端 X 是完全确定的。这时再令 $p(x_3)=p(x_5)=\frac{1}{2}$，检查它是否满足信道容量定理的条件，若满足，则该输入分布就是我们要求的最佳输入分布。可计算得

$$I(x_1;Y)=I(x_5;Y)=\log 2$$
$$I(x_2;Y)=I(x_4;Y)=0$$
$$I(x_3;Y)=0$$

满足信道容量定理的充要条件，因此信道容量

$$C=\log 2=1 \text{ 比特/符号}$$

若设 $p(x_1)=p(x_2)=p(x_4)=p(x_5)=\frac{1}{4}$，$p(x_3)=0$，也满足信道容量定理的充要条件，这时

$$\begin{aligned}I(x_1;Y)&=I(x_2;Y)\\&=I(x_4;Y)\\&=I(x_5;Y)\\&=\log 2\\I(x_3;Y)&<\log 2\end{aligned}$$

所以，该分布也是最佳分布。

可见，这个信道的最佳输入分布不是唯一的。由于 $I(x_i;Y)$ 仅直接与信道传递概率及输出符号概率有关，因此达到信道容量的输入概率分布不是唯一的，但输出概率分布是唯一的。

对于某些比较简单直观的信道，我们可以利用以上方法求信道容量。

3.2.7　信道容量的迭代算法*

前述几种方法都不能保证对于任意离散信道求出其信道容量。利用计算机的迭代算法可以以任意给定的精度及有限步数求出任意离散信道的信道容量。

$$\begin{aligned}I(X;Y)&=H(X)-H(X\mid Y)\\&=-\sum_i p(x_i)\ln p(x_i)+\sum_i\sum_j p(x_i)p(y_j\mid x_i)\ln p(x_i\mid y_j)\end{aligned}\tag{3.54}$$

对于某一固定的信道，其转移概率是已定的，所以 $I(X;Y)$ 是关于 $p(x_i)$ 和 $p(x_i \mid y_j)$ 的函数（上凸函数）。虽然事实上 $p(x_i \mid y_j)=\dfrac{p(x_i)p(y_j \mid x_i)}{\sum\limits_i p(x_i)p(y_j \mid x_i)}$ 也是 $p(x_i)$ 的函数，但可把 $I(X;Y)$ 视为关于 $p(x_i)$ 和 $p(x_i \mid y_j)$ 的函数，记为 $I[p(x_i),p(x_i \mid y_j)]$。

先固定变量 $p(x_i)$，求 $I[p(x_i),p(x_i \mid y_j)]$ 关于 $p(x_i \mid y_j)$ 的极值。这是在约束条件 $\sum\limits_i p(x_i \mid y_j)=1(j=1,2,\cdots,s)$ 下的条件极值。

利用拉格朗日乘子法，设辅助函数

$$F=I[p(x_i),p(x_i \mid y_j)]-\sum_j \lambda_j \sum_i p(x_i \mid y_j) \tag{3.55}$$

$$\begin{aligned}&\frac{\partial F}{\partial p(x_i \mid y_j)}\\&=\frac{\partial}{\partial p(x_i \mid y_j)}\left[-\sum_i p(x_i)\ln p(x_i)+\sum_i\sum_j p(x_i)p(y_j \mid x_i)\ln p(x_i \mid y_j)-\sum_j \lambda_j \sum_i p(x_i \mid y_j)\right]\\&=\frac{p(x_i)p(y_j \mid x_i)}{p(x_i \mid y_j)}-\lambda_j\end{aligned} \tag{3.56}$$

令 $\dfrac{\partial F}{\partial p(x_i \mid y_j)}=0$，则

$$\lambda_j=\frac{p(x_i)p(y_j \mid x_i)}{p(x_i \mid y_j)} \tag{3.57}$$

$$p(x_i \mid y_j)=\frac{p(x_i)p(y_j \mid x_i)}{\lambda_j} \tag{3.58}$$

其中，$i=1,2,\cdots,r$，$j=1,2,\cdots,s$。

利用约束条件 $\sum\limits_i p(x_i \mid y_j)=1$，得

$$\sum_i p(x_i \mid y_j)=\sum_i \frac{p(x_i)p(y_j \mid x_i)}{\lambda_j}=1 \tag{3.59}$$

所以

$$\lambda_j=\sum_i p(x_i)p(y_j \mid x_i) \qquad j=1,2,\cdots,s \tag{3.60}$$

因此，求得使 $I[p(x_i),p(x_i \mid y_j)]$ 达到极值的 $p(x_i \mid y_j)^*$ 为

$$p(x_i \mid y_j)^*=\frac{p(x_i)p(y_j \mid x_i)}{\sum\limits_i p(x_i)p(y_j \mid x_i)} \tag{3.61}$$

其中，$i=1,2,\cdots,r$，$j=1,2,\cdots,s$。

在求得 $p(x_i \mid y_j)^*$ 后，再固定 $p(x_i \mid y_j)$，求 $I[p(x_i),p(x_i \mid y_j)]$ 关于 $p(x_i)$ 的极值。此时的约束条件是 $\sum\limits_i p(x_i)=1$。

设辅助函数

$$\begin{aligned}Q&=I[p(x_i),p(x_i \mid y_j)]-\lambda\sum_i p(x_i)\\&=-\sum_i p(x_i)\ln p(x_i)+\sum_i\sum_j p(x_i)p(y_j \mid x_i)\ln p(x_i \mid y_j)-\lambda\sum_i p(x_i)\end{aligned} \tag{3.62}$$

$$\frac{\partial Q}{\partial p(x_i)}=-\ln p(x_i)-1+\sum_j p(y_j \mid x_i)\ln p(x_i \mid y_j)-\lambda \tag{3.63}$$

令$\frac{\partial Q}{\partial p(x_i)}=0$，得

$$-\ln p(x_i)-1+\sum_j p(y_j\mid x_i)\ln p(x_i\mid y_j)-\lambda=0 \tag{3.64}$$

$$p(x_i)=\exp\left[\sum_j p(y_j\mid x_i)\ln p(x_i\mid y_j)-\lambda-1\right]$$

$$=\frac{\exp\left[\sum_j p(y_j\mid x_i)\ln p(x_i\mid y_j)\right]}{\exp(1+\lambda)} \tag{3.65}$$

利用约束条件 $\sum_i p(x_i)=1$，得

$$\sum_i p(x_i)=\sum_i \frac{\exp\left[\sum_j p(y_j\mid x_i)\ln p(x_i\mid y_j)\right]}{\exp(1+\lambda)}=1 \tag{3.66}$$

$$\exp(1+\lambda)=\sum_i \exp\left[\sum_j p(y_j\mid x_i)\ln p(x_i\mid y_j)\right] \tag{3.67}$$

$$1+\lambda=\ln\sum_i \exp\left[\sum_j p(y_j\mid x_i)\ln p(x_i\mid y_j)\right] \tag{3.68}$$

所以，使 $I[p(x_i),p(x_i\mid y_j)]$ 达到极值的 $p(x_i)^*$ 为

$$p(x_i)^*=\frac{\exp\left[\sum_j p(y_j\mid x_i)\ln p(x_i\mid y_j)\right]}{\sum_i \exp\left[\sum_j p(y_j\mid x_i)\ln p(x_i\mid y_j)\right]} \tag{3.69}$$

其中，$i=1,2,\cdots,r$。移项得

$$-\ln p(x_i)+\sum_j p(y_j\mid x_i)\ln p(x_i\mid y_j)=1+\lambda \tag{3.70}$$

式(3.70)两端同乘以 $p(x_i)$，并对 i 求和，有

$$-\sum_i p(x_i)\ln p(x_i)+\sum_i p(x_i)\sum_j p(y_j\mid x_i)\ln p(x_i\mid y_j)=1+\lambda$$

得

$$I[p(x_i)^*,p(x_i\mid y_j)^*]=1+\lambda$$

$$=\ln\sum_i \exp\left[\sum_j p(y_j\mid x_i)\ln p(x_i\mid y_j)\right] \tag{3.71}$$

利用式(3.61)、式(3.69)、式(3.71)，便可以对信道容量进行迭代计算。

用迭代法计算信道容量的计算步骤如下：记 $p(y_j\mid x_i)=p_{ij}$，$p(x_i)=p_i$，$p(x_i\mid y_j)=\varphi_{ji}$ $(i=1,2,\cdots,r;\ j=1,2,\cdots,s)$。

算法：

① 初始化信源分布 $\boldsymbol{p}^{(0)}=(p_1,p_2,\cdots,p_i,\cdots,p_r)$（一般初始化为均匀分布），置迭代计数器 $k=0$，设信道容量相对误差门限 $\delta(\delta>0)$。

② $$\varphi_{ji}^{(k)}=\frac{p_{ij}p_i^{(k)}}{\sum_i p_{ij}p_i^{(k)}} \tag{3.72}$$

其中，$i=1,2,\cdots,r;j=1,2,\cdots,s$。

③ $$p_i^{(k+1)}=\frac{\exp\left[\sum_j p_{ij}\ln\varphi_{ji}^{(k)}\right]}{\sum_i\left\{\exp\left[\sum_j p_{ij}\ln\varphi_{ji}^{(k)}\right]\right\}} \tag{3.73}$$

其中，$i=1,2,\cdots,r$。

④ $$C^{(k+1)}=\ln\left\{\sum_i \exp\left[\sum_j p_{ij}\ln\varphi_{ji}^{(k)}\right]\right\} \tag{3.74}$$

⑤ 如果 $\Delta C^{(k+1)}=\dfrac{|C^{(k+1)}-C^{(k)}|}{C^{(k+1)}}\leqslant\delta$，转向⑦。

⑥ 置迭代序号 $k+1\to k$，转向②。

⑦ 输出 $p_i^{(k+1)}$ 和 $C^{(k+1)}$ 的结果。

⑧ 停止。

可以证明，平均互信息 $I[p(x_i),p(x_i\mid y_j)]$ 具有收敛性，即 $\lim\limits_{k\to\infty}|C^{(k+1)}-C^{(k)}|=0$，所以迭代算法最终能求出任意精度的解。算法的收敛速度与信源初始概率分布的选择有很大关系，若初始分布选得越接近最佳输入分布，则收敛的速度越快，若初始分布选得正好是最佳输入分布，则一步就求得信道容量。

3.3 离散多符号信道及其信道容量

3.2 节中讨论了最简单的离散信道，即信道的输入和输出都只是单个随机变量的信道。实际离散信道的输入和输出常常是随机变量序列，用随机矢量来表示，称为离散多符号信道（如图 3.8 所示）。实际离散信道往往是有记忆信道。为简化起见，我们主要研究离散无记忆信道。

图 3.8 离散多符号信道模型

【定义 3-6】若在任意时刻信道的输出只与此时刻信道的输入有关，而与其他时刻的输入和输出无关，则称之为离散无记忆信道（Discrete Memoryless Channel，DMC）。

输入、输出随机序列的长度为 N 的离散无记忆平稳信道，通常称为离散无记忆信道的 N 次扩展信道。

输入随机序列 $\boldsymbol{X}=X_1X_2\cdots X_N$ 中，随机变量 $X_i(i=1,2,\cdots,N)$ 都取值于同一输入符号集 X，而符号集 X 有 r 个符号，所以随机矢量 X 的可能取值有 r^N 个。输出随机序列 $\boldsymbol{Y}=Y_1Y_2\cdots Y_N$ 中，随机变量 $Y_i(i=1,2,\cdots,N)$ 都取值于同一输入符号集 Y，而符号集 Y 有 s 个符号，所以随机矢量 $\boldsymbol{Y}$ 的可能取值共有 s^N 个，因此 N 次扩展信道的信道矩阵是一个 $r^N\times s^N$ 矩阵。离散无记忆信道的数学模型仍然表示为 $\{\boldsymbol{X},P(\boldsymbol{Y}\mid\boldsymbol{X}),\boldsymbol{Y}\}$。这时的输入、输出均为随机矢量。

根据信道无记忆的特性，其转移概率

$$\begin{aligned}P(\boldsymbol{Y}\mid\boldsymbol{X})&=P(Y_1Y_2\cdots Y_N\mid X_1X_2\cdots X_N)\\&=P(Y_1\mid X_1)P(Y_2\mid X_2)\cdots P(Y_N\mid X_N)\\&=\prod_{k=1}^{N}P(Y_k\mid X_k)\end{aligned} \tag{3.75}$$

【例 3.8】求二元对称信道的二次扩展信道的信道矩阵。

【解】二元对称信道的二次扩展信道的输入、输出序列的每个随机变量均取值于$\{0,1\}$，输入共有$r^N=2^2=4$个取值，输出共有$s^N=2^2=4$个取值。根据

$$P(\boldsymbol{Y}\mid\boldsymbol{X})=\prod_{k=1}^{N}P(Y_k\mid X_k)$$

可求出

$$\begin{aligned}
p(\boldsymbol{y}_1\mid\boldsymbol{x}_1)&=p(00\mid00)=p(0\mid0)p(0\mid0)=\bar{p}^2\\
p(\boldsymbol{y}_2\mid\boldsymbol{x}_1)&=p(01\mid00)=p(0\mid0)p(1\mid0)=\bar{p}p\\
p(\boldsymbol{y}_3\mid\boldsymbol{x}_1)&=p(10\mid00)=p(1\mid0)p(0\mid0)=p\bar{p}\\
p(\boldsymbol{y}_4\mid\boldsymbol{x}_1)&=p(11\mid00)=p(1\mid0)p(1\mid0)=p^2
\end{aligned}$$

同理可求出其他转移概率$p_{ij}(i=1,2,3,4,\ j=1,2,3,4)$，得到信道矩阵：

$$\boldsymbol{P}=\begin{bmatrix}
\bar{p}^2 & \bar{p}p & p\bar{p} & p^2\\
\bar{p}p & \bar{p}^2 & p^2 & p\bar{p}\\
\bar{p}\bar{p} & p^2 & \bar{p}^2 & \bar{p}p\\
p^2 & p\bar{p} & \bar{p}p & \bar{p}^2
\end{bmatrix}$$

二元对称信道的二次扩展信道如图 3.9 所示。

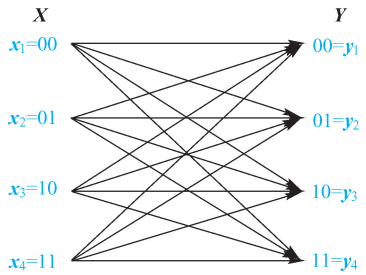


图 3.9　二元对称信道的二次扩展信道

关于离散无记忆信道的平均互信息有以下定理。

【定理 3-4】若信道的输入和输出分别是 N 长序列 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$，且信道是无记忆的，则

$$I(\boldsymbol{X};\boldsymbol{Y})\leqslant\sum_{k=1}^{N}I(X_k;Y_k)\tag{3.76}$$

其中，X_k、Y_k分别是序列 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 中的第 k 位随机变量。

【证明】

$$I(\boldsymbol{X};\boldsymbol{Y})=H(\boldsymbol{Y})-H(\boldsymbol{Y}\mid\boldsymbol{X})\tag{3.77}$$

根据熵函数的链规则以及条件熵和无条件熵的关系，可得

$$\begin{aligned}
H(\boldsymbol{Y})&=H(Y_1Y_2\cdots Y_N)\\
&=H(Y_1)+H(Y_2\mid Y_1)+\cdots+H(Y_N\mid Y_1Y_2\cdots Y_{N-1})\\
&\leqslant\sum_{k=1}^{N}H(Y_k)
\end{aligned}\tag{3.78}$$

根据熵函数的链规则以及离散无记忆信道的定义，可得

$$
\begin{aligned}
H(\boldsymbol{Y}\mid\boldsymbol{X}) &= H(Y_1Y_2\cdots Y_N\mid X_1X_2\cdots X_N)\\
&= H(Y_1\mid X_1X_2\cdots X_N)+H(Y_2\mid X_1X_2\cdots X_NY_1)+\cdots+\\
&\quad H(Y_N\mid X_1X_2\cdots X_NY_1Y_2\cdots Y_{N-1})\\
&= \sum_{k=1}^{N}H(Y_k\mid X_k)
\end{aligned}
\tag{3.79}
$$

所以

$$
\begin{aligned}
I(\boldsymbol{X};\boldsymbol{Y}) &\leqslant \sum_{k=1}^{N}H(Y_k)-\sum_{k=1}^{N}H(Y_k\mid X_k)\\
&= \sum_{k=1}^{N}I(X_k;Y_k)
\end{aligned}
$$

即对于离散无记忆信道，其平均互信息 $I(\boldsymbol{X};\boldsymbol{Y})$ 小于等于序列 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 中所有对应时刻的单个随机变量 X_k、Y_k 的平均互信息 $I(X_k;Y_k)$ 之和。当且仅当信源也是无记忆信源时，等号成立。

当信源是无记忆信源时，有

$$
\begin{aligned}
P(\boldsymbol{X}) &= \prod_{k=1}^{N}P(X_k)\\
P(\boldsymbol{XY}) &= P(\boldsymbol{X})P(\boldsymbol{Y}\mid\boldsymbol{X})\\
&= \prod_{k=1}^{N}P(X_k)\prod_{k=1}^{N}P(Y_k\mid X_k)\\
&= \prod_{k=1}^{N}P(X_k)P(Y_k\mid X_k)\\
&= \prod_{k=1}^{N}P(X_kY_k)
\end{aligned}
\tag{3.80}
$$

$$
\begin{aligned}
p(\boldsymbol{y}_j) &= \sum_{i=1}^{r^N}p(\boldsymbol{x}_i\boldsymbol{y}_j)\\
&= \sum_{i_1=1}^{r}p(x_{i_1}y_{j_1})\sum_{i_2=1}^{r}p(x_{i_2}y_{j_2})\cdots\sum_{i_N=1}^{r}p(x_{i_N}y_{j_N})\\
&= \prod_{k=1}^{N}p(y_{j_k})
\end{aligned}
\tag{3.81}
$$

则

$$
P(\boldsymbol{Y}) = \prod_{k=1}^{N}P(Y_k)
$$

$$
H(\boldsymbol{Y}) = \sum_{k=1}^{N}H(Y_k) \tag{3.82}
$$

$$
I(\boldsymbol{X};\boldsymbol{Y}) = \sum_{k=1}^{N}I(X_k;Y_k) \tag{3.83}
$$

即信源和信道均为无记忆时，其序列 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的平均互信息 $I(\boldsymbol{X};\boldsymbol{Y})$ 等于序列中所有对应时刻单个随机变量 X_k、Y_k 的平均互信息 $I(X_k;Y_k)$ 之和。

证毕。

如果信道输入序列中的每个随机变量均取值于同一信源符号集并且具有同一种概率分布(取自于同一概率空间)，通过相同的信道传到输出端，则输出序列中的每个随机变量也取自同一符号集，并且具有相同的概率分布，因此有

$$\begin{aligned} X_1 &= X_2 = \cdots = X_N = X \\ Y_1 &= Y_2 = \cdots = Y_N = Y \\ I(X_1;Y_1) &= I(X_2;Y_2) = \cdots = I(X_N;Y_N) = I(X;Y) \end{aligned} \tag{3.84}$$

于是

$$\begin{aligned} I(\boldsymbol{X};\boldsymbol{Y}) &= \sum_{k=1}^{N} I(X_k;Y_k) \\ &= NI(X;Y) \end{aligned} \tag{3.85}$$

式(3.85)表明，对于离散无记忆 N 次扩展信道，当信源是平稳无记忆信源时，其平均互信息 $I(\boldsymbol{X};\boldsymbol{Y})$ 等于单符号信道的平均互信息的 N 倍。

离散无记忆信道的 N 次扩展信道的信道容量为

$$\begin{aligned} C^N &= \max_{P(\boldsymbol{X})} I(\boldsymbol{X};\boldsymbol{Y}) \\ &= \max_{P(\boldsymbol{X})} \sum_{k=1}^{N} I(X_k;Y_k) \\ &= \sum_{k=1}^{N} \max_{P(X_k)} I(X_k;Y_k) \\ &= \sum_{k=1}^{N} C_k \end{aligned} \tag{3.86}$$

其中，$C_k = \max\limits_{P(X_k)} I(X_k;Y_k)$ 是时刻 k 通过离散无记忆信道传输的最大信息量，可以用前面介绍的求离散单符号信道的信道容量的方法求解。

因为现在输入随机序列 $\boldsymbol{X} = (X_1, \cdots, X_k, \cdots, X_N)$ 在同一信道中传输，所以任何时刻通过离散无记忆信道传输的最大信息量都相同，即 $C_k = C(k = 1,2,\cdots,N)$，则

$$C^N = NC \tag{3.87}$$

离散无记忆信道的 N 次扩展信道的信道容量等于原单符号离散信道的信道容量的 N 倍，当信源也是无记忆信源并且每一时刻的输入分布各自达到最佳输入分布时，才能达到这个信道容量 NC。

一般，消息序列在离散无记忆 N 次扩展信道中传输时，平均互信息量 $I(\boldsymbol{X};\boldsymbol{Y}) \leqslant NC$。

3.4 组合信道及其信道容量

前面分析了单符号离散信道和离散无记忆信道的扩展信道。实际应用中常常会遇到两个或更多个信道组合在一起使用的情况。例如，待发送的消息比较多时，可能要用两个或更多个信道并行发送，这种组合信道称为并联信道；有时消息会依次通过几个信道串联发送，如无线电中继信道、数据处理系统，这种组合信道称为级联信道。在研究比较复杂的信道时，

为使问题简化，往往可以将它们分解成几个简单的信道的组合。本节将讨论这两种组合信道的信道容量与其组成信道的信道容量之间的关系。

3.4.1 独立并联信道

一般的独立并联信道如图 3.10 所示。

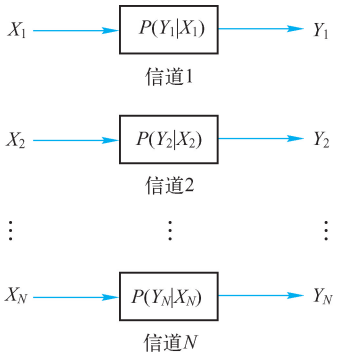


图 3.10 独立并联信道

设有 N 个信道并联，它们的输入分别为 $X_1, X_2, \cdots, X_N$，输出分别是 $Y_1, Y_2, \cdots, Y_N$，N 个信道的传递概率分别是 $P(Y_1 \mid X_1), P(Y_2 \mid X_2), \cdots, P(Y_N \mid X_N)$。在这 N 个独立信道中，每个信道的输出 Y_k 只与本信道的输入 X_k 有关，而与其他信道的输入、输出无关。这 N 个信道的联合传递概率满足以下关系：

$$P(Y_1Y_2\cdots Y_N \mid X_1X_2\cdots X_N) = P(Y_1 \mid X_1)P(Y_2 \mid X_2)\cdots P(Y_N \mid X_N) \tag{3.88}$$

这相当于信道是无记忆时应满足的条件。因此我们可以把定理 3-4 的结论推广到 N 个独立并联信道中：

$$I(X_1X_2\cdots X_N; Y_1Y_2\cdots Y_N) \leqslant \sum_{k=1}^{N} I(X_k; Y_k)$$

即联合平均互信息不大于各信道的平均互信息之和。因此，独立并联信道的信道容量

$$C_{并} = \max_{P(X_1\cdots X_N)} I(X_1\cdots X_N; Y_1\cdots Y_N) \leqslant \sum_{k=1}^{N} C_k \tag{3.89}$$

其中，$C_k = \max_{P(X_k)} I(X_k; Y_k)$ 是各独立信道的信道容量。

所以，独立并联信道的信道容量不大于各信道的信道容量之和。只有当每个输入随机变量的概率分布均达到各自信道的最佳输入分布时，独立并联信道的信道容量才等于各信道容量之和，即

$$C_{并} = \sum_{k=1}^{N} C_k \tag{3.90}$$

当 N 个独立并联信道的信道容量都相同时，则

$$C_{并} = NC \tag{3.91}$$

3.4.2 级联信道

级联信道是信道最基本的组合形式，许多实际信道都可视为其组成信道的级联。

图 3.11 是由两个单符号信道组成的最简单的级联信道。

信道 I 的输入随机变量为 X，输出随机变量为 Y。信道 II 的输入随机变量为 Y，输出随机变量为 Z。信道 I 的输出恰好是信道 II 的输入。信道 I 的输出 Y 与输入 X 统计相关，而信道 II 的输出 Z 与输入 Y 统计相关，一般来说，Z 将与 X 统计相关。但是级联的结构又决定了在给定 Y 以后，Z 的取值将不再与 X 有关，而只取决于信道 II 的前向转移概率 $P(Z|Y)$，也就是说，$X\to Y\to Z$ 组成一个马尔可夫链。根据马尔可夫链的性质，级联信道的总信道矩阵等于这两个串接信道的信道矩阵的乘积。求得级联信道的总信道矩阵后，级联信道的信道容量就可以用求离散单符号信道的信道容量的方法计算。

【例 3.9】设有两个离散二元对称信道，其级联信道如图 3.12 所示，求级联信道的信道容量。

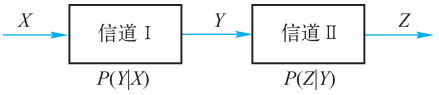


图 3.11　级联信道

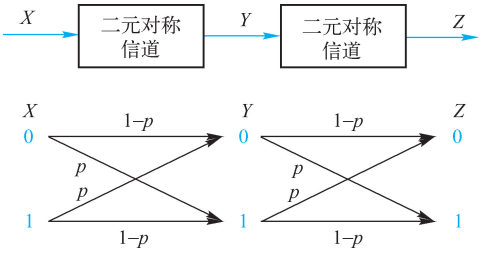


图 3.12　二元对称信道的级联信道

【解】

两个二元对称信道的信道矩阵均为

$$\boldsymbol{P}_1=\boldsymbol{P}_2=\begin{bmatrix}1-p & p\\ p & 1-p\end{bmatrix}$$

因为 X、Y、Z 组成马尔可夫链，则级联信道的总信道矩阵为

$$\begin{aligned}\boldsymbol{P}&=\boldsymbol{P}_1\boldsymbol{P}_2\\&=\begin{bmatrix}1-p & p\\ p & 1-p\end{bmatrix}\begin{bmatrix}1-p & p\\ p & 1-p\end{bmatrix}\\&=\begin{bmatrix}(1-p)^2+p^2 & 2p(1-p)\\ 2p(1-p) & (1-p)^2+p^2\end{bmatrix}\end{aligned}$$

所以，级联信道仍然是一个二元对称信道，即

$$C_{\text{串}}=1-H[2p(1-p)]$$

3.5　连续信道及其信道容量

3.5.1　连续随机变量的互信息

连续随机变量 X 和 Y 之间的平均互信息定义为

$$I(X;Y)=\iint_{\mathbf{R}^2}p(xy)\log\frac{p(xy)}{p(x)p(y)}\mathrm{d}x\mathrm{d}y \tag{3.92}$$

连续随机变量的平均互信息 $I(X;Y)$ 的计算和离散随机变量一样，只要将离散情况下的概率分布换成概率密度，求和化成积分即可。连续随机变量的平均互信息具有和离散随机变量的平均互信息一样的性质：

① 对称性

$$\begin{aligned} I(X;Y) &= I(Y;X) \\ &= h(X) - h(X \mid Y) \\ &= h(Y) - h(Y \mid X) \\ &= h(X) + h(Y) - h(XY) \end{aligned} \tag{3.93}$$

② 非负性

$$I(X;Y) \geqslant 0 \tag{3.94}$$

当且仅当随机变量 X 和 Y 统计独立时，等号成立。

因此，虽然连续随机变量的熵不具有非负性，但连续随机变量的熵差 $I(X;Y)$ 仍具有非负性。

【例 3.10】设 $p(xy)$ 是二维高斯随机变量 XY 的概率密度函数

$$p(xy) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-m_X)^2}{\sigma_X^2} - \frac{2\rho(x-m_X)(y-m_Y)}{\sigma_X\sigma_Y} + \frac{(y-m_Y)^2}{\sigma_Y^2}\right]\right\}$$

求 $I(X;Y)$。其中，m_X、m_Y、σ_X^2、σ_Y^2 分别表示随机变量 X 和 Y 的均值和方差，ρ 是归一化相关函数

$$\rho = \frac{E[(X-E(X))(Y-E(Y))]}{\sigma_X\sigma_Y}$$

【解】

（1）先求 X 和 Y 的一维概率密度函数：

$$\begin{aligned} p(x) &= \int_{-\infty}^{+\infty} p(xy)\,\mathrm{d}y \\ &= \frac{1}{\sqrt{2\pi}\sigma_X}\exp\left[-\frac{(x-m_X)^2}{2\sigma_X^2}\right] \end{aligned}$$

$$\begin{aligned} p(y) &= \int_{-\infty}^{+\infty} p(xy)\,\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}\sigma_Y}\exp\left[-\frac{(y-m_Y)^2}{2\sigma_Y^2}\right] \end{aligned}$$

（2）由平均互信息的定义求得

$$\begin{aligned} I(X;Y) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} p(xy)\ln\frac{p(xy)}{p(x)p(y)}\mathrm{d}x\mathrm{d}y \\ &= \ln\frac{1}{\sqrt{1-\rho^2}} - \frac{1}{2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left[\frac{(x-m_X)^2}{(1-\rho^2)\sigma_X^2} - \frac{2\rho(x-m_X)(y-m_Y)}{(1-\rho^2)\sigma_X\sigma_Y} + \right. \\ &\quad \left.\frac{(y-m_Y)^2}{(1-\rho^2)\sigma_Y^2} - \frac{(x-m_X)^2}{\sigma_X^2} - \frac{(y-m_Y)^2}{\sigma_Y^2}\right]p(xy)\,\mathrm{d}x\mathrm{d}y \end{aligned}$$

$$= -\frac{1}{2}\ln(1-\rho^2) - \frac{1}{2}\left[\frac{1}{1-\rho^2} - \frac{2\rho^2}{1-\rho^2} + \frac{1}{1-\rho^2} - 1 - 1\right]$$

$$= -\frac{1}{2}\ln(1-\rho^2) \text{ 奈特/样值}$$

以上表明两个高斯随机变量之间的互信息只与相关系数 ρ 有关，与数学期望 m_X、m_Y 及方差 σ_X^2、σ_Y^2 无关。数学期望 m_X、m_Y 代表变量的直流成分，而直流成分不会含有任何信息。互信息只与归一化相关函数值或功率的相对大小有关，与功率的绝对大小无关。这与我们的经验是一致的。

3.5.2 加性高斯信道的信道容量

可以证明，连续信道输入、输出随机变量的平均互信息，是信源的概率密度 $p(x)$ 的上凸函数。我们仍然定义连续信道的信道容量为平均互信息关于信源概率密度函数的极大值，即

$$C = \max_{p(x)} I(X;Y) \tag{3.95}$$

一般来说，连续信道的信道容量并不容易计算，加性噪声信道则相对简单。下面只研究这种信道，即噪声（记为连续随机变量 N）与输入随机变量 X 相互统计独立。这种信道噪声对输入的干扰作用表现为输出是噪声和输入的线性叠加，即 $Y = X + N$。可以证明，当噪声功率 σ_N^2 给定后，高斯噪声的信道容量 C 最小，因此高斯信道是最差的信道。实际应用中往往把噪声视为高斯噪声。

对于加性噪声信道，由坐标变换理论可以证明 $p(y \mid x) = p(n)$，其中 $p(n)$ 是噪声 N 的概率密度函数，也就是说信道的条件概率密度函数等于噪声的概率密度函数。这时

$$\begin{aligned} h(Y \mid X) &= -\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} p(xy)\log p(y \mid x)\,\mathrm{d}x\mathrm{d}y \\ &= -\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} p(x)p(y \mid x)\log p(y \mid x)\,\mathrm{d}x\mathrm{d}y \\ &= -\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} p(x)p(n)\log p(n)\,\mathrm{d}x\mathrm{d}n \\ &= -\int_{-\infty}^{+\infty} p(n)\log p(n)\,\mathrm{d}n \\ &= h(N) \end{aligned} \tag{3.96}$$

该结论进一步说明，条件熵 $h(Y \mid X)$ 是由信道中的噪声引起的，它完全等于噪声的信源熵，所以称为噪声熵。

下面研究噪声源为高斯白噪声的加性信道，其信道容量为

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) \\ &= \max_{p(x)} [h(Y) - h(Y \mid X)] \\ &= \max_{p(x)} [h(Y) - h(N)] \end{aligned} \tag{3.97}$$

由于加性信道的噪声 N 和信源 X 相互统计独立，X 的概率密度 $p(x)$ 的变动不会引起噪声熵 $h(N)$ 的改变，所以通过选择 $p(x)$ 使输出随机变量熵 $h(Y)$ 达到最大值时，加性信道即达到信道容量：

$$C=\max_{p(x)}h(Y)-h(N) \tag{3.98}$$

对于不同的限制条件，连续随机变量具有不同的最大值，所以连续信道的信道容量取决于输入随机变量 X 所受的限制条件以及噪声 N（即信道）的统计特性。

如果噪声 N 是均值为 0、方差为 σ_N^2 的高斯随机变量，即满足

$$\int_{-\infty}^{+\infty}p(n)\,\mathrm{d}n=1$$

$$\int_{-\infty}^{+\infty}np(n)\,\mathrm{d}n=0$$

$$\int_{-\infty}^{+\infty}n^2p(n)\,\mathrm{d}n=\sigma_N^2=P_N$$

其中，P_N 表示噪声 N 的平均功率，那么这种信道称为高斯加性连续信道。

一般来说，输入随机变量 X 的平均功率是有限的，假设限定为 P_X，而噪声的平均功率限定为 $P_N=\sigma_N^2$，则输出随机变量 Y 的平均功率也是有限的，设为 P_Y。根据最大连续熵定理，要使 $h(Y)$ 达到最大，Y 必须是一个高斯随机变量。而当输入 $p(x)$ 满足什么条件时才能使 Y 为高斯分布呢？

由概率论的知识可知，当 X、N 统计独立且 $Y=X+N$ 时，若输入 X 是均值为 0、方差为 $\sigma_X^2=P_X$ 的高斯随机变量，即

$$p(x)=\frac{1}{\sqrt{2\pi\sigma_X^2}}\mathrm{e}^{-\frac{x^2}{2\sigma_X^2}}$$

则

$$p(y)=\frac{1}{\sqrt{2\pi\sigma_Y^2}}\mathrm{e}^{-\frac{y^2}{2\sigma_Y^2}}$$

Y 为高斯分布，并且

$$\sigma_Y^2=\sigma_X^2+\sigma_N^2=P_Y \tag{3.99}$$

也就是说，当输入随机变量 X 的概率密度是均值为 0、方差为 σ_X^2 的高斯随机变量，加性信道的噪声 N 是均值为 0、方差为 σ_N^2 的高斯随机变量时，输出随机变量 Y 也是一个高斯随机变量，其均值为 0、方差为 $\sigma_Y^2=\sigma_X^2+\sigma_N^2=P_Y$，此时输出随机变量的熵 $h(Y)$ 达到最大，而信道达到信道容量：

$$\begin{aligned}C&=\max_{p(x)}h(Y)-h(N)\\&=\frac{1}{2}\log 2\pi\mathrm{e}(\sigma_X^2+\sigma_N^2)-\frac{1}{2}\log 2\pi\mathrm{e}\sigma_N^2\\&=\frac{1}{2}\log\frac{\sigma_X^2+\sigma_N^2}{\sigma_N^2}=\frac{1}{2}\log\left(1+\frac{\sigma_X^2}{\sigma_N^2}\right)\\&=\frac{1}{2}\log\left(1+\frac{P_X}{P_N}\right)\end{aligned} \tag{3.100}$$

其中，$\dfrac{P_X}{P_N}$称为信道的信噪比。

3.5.3 多维高斯加性信道的信道容量

当信道为多维高斯加性信道时，由于加性噪声信道必然是一个无记忆信道，所以

$$I(\boldsymbol{X};\boldsymbol{Y}) \leqslant \sum_{i=1}^{n} I(X_i;Y_i)$$

$$\leqslant \frac{1}{2}\sum_{i=1}^{n}\log\left(1+\frac{P_{X_i}}{P_{N_i}}\right) \tag{3.101}$$

因此

$$C=\max_{p(\boldsymbol{X})} I(\boldsymbol{X};\boldsymbol{Y})$$

$$=\max_{p(\boldsymbol{X})}\sum_{i=1}^{n} I(X_i;Y_i)$$

$$=\frac{n}{2}\log\left(1+\frac{P_{X_i}}{P_{N_i}}\right) \qquad i=1,2,\cdots,N \tag{3.102}$$

当且仅当输入随机矢量 $\boldsymbol{X}$ 中各分量统计独立并且均为高斯变量时，达到信道容量。

如果在每个抽样时刻信源和噪声是均值为 0、方差分别为 σ_X^2 和 σ_N^2 的高斯随机变量，则

$$C=\frac{n}{2}\log\left(1+\frac{\sigma_X^2}{\sigma_N^2}\right)\text{比特/}n\text{ 个样值} \tag{3.103}$$

3.6 波形信道的信道容量

波形信道通常根据抽样定理转化成多维连续信道进行处理。

一般来说，信道的带宽总是有限的。假设某信道的频带限于$(0,B)$，则其输入、输出信号和噪声都是限频的随机过程，频带限于$(0,B)$。根据抽样定理，可把一个时间连续的信道变换成时间离散的随机序列信道来处理，即用每隔$\frac{1}{2B}$秒时间的采样值来表示输入、输出信号和噪声。我们把一次采样视为信道的一次传输，由于每秒传输 $2B$ 个样值，所以单位时间的信道容量为

$$C_t=B\log\left(1+\frac{\sigma_X^2}{\sigma_N^2}\right)\text{比特/秒} \tag{3.104}$$

当噪声是双边功率谱密度为$\frac{N_0}{2}$的高斯白噪声时，有

$$C_t=B\log\left(1+\frac{\sigma_X^2}{N_0B}\right) \tag{3.105}$$

这就是著名的香农公式，适用于加性高斯白噪声信道。从前面的讨论可知，只有当输入信号为功率受限的高斯白噪声信号时，才能达到该信道容量。

香农公式说明，当信道容量一定时，增大信道的带宽，可以降低对信噪功率比的要求；反之，当信道频带较窄时，可以通过提高信噪功率比来补偿。

当 $B\to\infty$ 时，有

$$C_t=\lim_{B\to\infty}B\log\left(1+\frac{\sigma_X^2}{N_0B}\right)$$

$$=\lim_{B\to\infty}\frac{\sigma_X^2}{N_0}\times\frac{N_0B}{\sigma_X^2}\log\left(1+\frac{\sigma_X^2}{N_0B}\right)$$

$$= \frac{\sigma_X^2}{N_0} \times \log e$$

$$= 1.44 \frac{\sigma_X^2}{N_0} \tag{3.106}$$

上式表明，当频带很宽时，信道容量正比于信号功率与噪声谱密度之比，是加性高斯噪声信道信息传输率的极限值。

【例 3.11】一般模拟电话信道的带宽为 3300 Hz，若信噪比为 20 dB$\left(即\frac{\sigma_X^2}{N_0 B}=100\right)$，则根据香农公式可得电话信道的信道容量：

$$C_t = B\log\left(1+\frac{\sigma_X^2}{N_0 B}\right) = 22000 \text{ 比特/秒}$$

由于高斯加性信道是实际信道中最差的信道，所以香农公式可用于确定实际信道的信道容量的下限值。香农公式给出了在有噪信道中无失真传输所能达到的极限信息传输率，因此对实际通信系统有非常重要的指导意义。

扩展阅读：信道容量定理引理

【引理 3-1】

$$\lim_{\theta \to 0} \frac{1}{\theta} \{ I[\theta \boldsymbol{q} + (1-\theta)\boldsymbol{p}] - I(\boldsymbol{p}) \} = \sum_{i=1}^{r} (q_i - p_i) \frac{\partial I(\boldsymbol{p})}{\partial p_i}$$

这里把 $I(X;Y)$ 写成概率矢量 $\boldsymbol{p}$、$\boldsymbol{q}$ 的函数 $I(\boldsymbol{p})$ 和 $I(\boldsymbol{q})$ 的形式，$\boldsymbol{p}$、$\boldsymbol{q}$ 为不同的概率矢量：

$$\boldsymbol{p} = (p_1, p_2, \cdots, p_r)$$

$$\boldsymbol{q} = (q_1, q_2, \cdots, q_r)$$

【证明】

$$\begin{aligned}
&I[\theta \boldsymbol{q} + (1-\theta)\boldsymbol{p}] - I(\boldsymbol{p}) \\
&= I[\boldsymbol{p} + \theta(\boldsymbol{q} - \boldsymbol{p})]] - I(\boldsymbol{p}) \\
&= I[p_1 + \theta(q_1 - p_1), p_2 + \theta(q_2 - p_2), p_3 + \theta(q_3 - p_3), \cdots, p_r + \theta(q_r - p_r)] - \\
&\quad I(p_1, p_2, \cdots, p_r) \\
&= I[p_1 + \theta(q_1 - p_1), p_2 + \theta(q_2 - p_2), p_3 + \theta(q_3 - p_3), \cdots, p_r + \theta(q_r - p_r)] - \\
&\quad I[p_1, p_2 + \theta(q_2 - p_2), p_3 + \theta(q_3 - p_3), \cdots, p_r + \theta(q_r - p_r)] + \\
&\quad I[p_1, p_2 + \theta(q_2 - p_2), p_3 + \theta(q_3 - p_3), \cdots, p_r + \theta(q_r - p_r)] + \\
&\quad I[p_1, p_2, p_3 + \theta(q_3 - p_3), \cdots, p_r + \theta(q_r - p_r)] - \\
&\quad I[p_1, p_2, p_3 + \theta(q_3 - p_3), \cdots, p_r + \theta(q_r - p_r)] \\
&\qquad\qquad \vdots \\
&= I[p_1, p_2, p_3, \cdots, p_r + \theta(q_r - p_r)] - I(p_1, p_2, \cdots, p_r)
\end{aligned}$$

上式为 r 对 r 个分量的多元函数的减式，每对中只有一个分量不同，其余分量相同。对每对减式除以 θ 并求 $\theta\to 0$ 时的极限，这可以视为多元函数对该分量求偏导，即

$$\lim_{\theta\to 0}\frac{1}{\theta}I[p_1,p_2,\cdots,p_i+\theta(q_i-p_i),\cdots,p_r]-I[p_1,p_2,\cdots,p_i,\cdots,p_r]=(q_i-p_i)\frac{\partial I(\boldsymbol{p})}{\partial p_i}$$

所以

$$\begin{aligned}&\lim_{\theta\to 0}\frac{1}{\theta}\{I[\theta\boldsymbol{q}+(1-\theta)\boldsymbol{p}]-I(\boldsymbol{p})\}\\&=(q_1-p_1)\frac{\partial I(\boldsymbol{p})}{\partial p_1}+(q_2-p_2)\frac{\partial I(\boldsymbol{p})}{\partial p_2}+\cdots+\\&\quad(q_r-p_r)\frac{\partial I(\boldsymbol{p})}{\partial p_r}\\&=\sum_{i=1}^{r}(q_i-p_i)\frac{\partial I(\boldsymbol{p})}{\partial p_i}\end{aligned}$$

动手实践：信道容量的迭代算法

【已知】信源符号个数 r、信宿符号个数 s、信道转移概率矩阵 $\boldsymbol{P}=(p_{ji})^{r\times s}$。

【算法】

① 初始化信源分布 $p_i=\frac{1}{r}$，循环变量 $k=1$，门限 Δ，$C^{(0)}=-\infty$。

② $$\phi_{ij}^{(k)}=\frac{p_i^{(k)}p_{ji}}{\sum_{i=1}^{r}p_i^{(k)}p_{ji}}$$

③ $$p_i^{(k+1)}=\frac{\exp\left(\sum_{j=1}^{s}p_{ji}\ln\phi_{ij}^{(k)}\right)}{\sum_{i=1}^{r}\exp\left(\sum_{j=1}^{s}p_{ji}\ln\phi_{ij}^{(k)}\right)}$$

④ $$C^{(k+1)}=\log\left[\sum_{i=1}^{r}\exp\left(\sum_{j=1}^{s}p_{ji}\ln\phi_{ij}^{(k)}\right)\right]$$

⑤ 若 $\frac{|C^{(k+1)}-C^{(k)}|}{C^{(k+1)}}>\Delta$，则 $k=k+1$，转第②步。

⑥ 输出 $\overline{P}^{*}=p_i^{(k+1)}$ 和 $C^{(k+1)}$，终止。

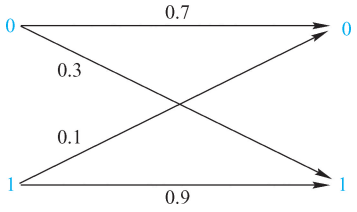
【要求】

（1）输入：任意的一个信道转移概率矩阵。信源符号个数、信宿符号个数和每个具体的转移概率在运行时从键盘输入。

（2）输出：最佳信源分布 $\overline{P}^{*}$，信道容量 C。

习　题　3

3.1　一个二元信道如题图 3.1 所示，其输入概率空间为$\begin{bmatrix} X \\ P \end{bmatrix}=\begin{bmatrix} 0 & 1 \\ 0.2 & 0.8 \end{bmatrix}$，试计算 $I(x=0;y=1)$，$I(x=1;Y)$ 和 $I(X;Y)$。



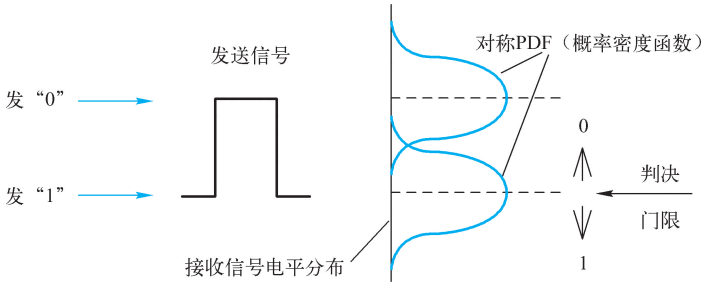
题图 3.1

3.2　二元删除信道有两个输入(0,1)和三个输出(0,1,E)，其中 E 表示可检出但无法纠正的错误。信道前向转移概率是

$$p(0|0)=1-\alpha \qquad p(E|0)=\alpha \qquad p(1|0)=0$$
$$p(0|1)=0 \qquad p(E|1)=\alpha \qquad p(1|1)=1-\alpha$$

求信道容量 C。

3.3　设某二进制数字传输系统接收判决器的输入信号电平、噪声密度分布、判决电平如题图 3.3 所示。试求：（1）信道模型；（2）互信息；（3）信道容量。

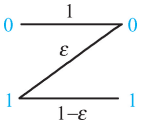


题图 3.3

3.4　设有干扰离散信道的输入端是以等概率出现的 A、B、C、D 四个字母。该信道的正确传输概率为$\frac{1}{2}$，错误传输概率平均分布在其他三个字母上。验证：在该信道上每个字母传输的平均信息量为 0.21 比特。

3.5　Z 信道及其输入、输出如题图 3.5 所示。

$$p(x|y)=\begin{bmatrix} 1 & 0 \\ \varepsilon & 1-\varepsilon \end{bmatrix} \qquad x,y\in\{0,1\}$$



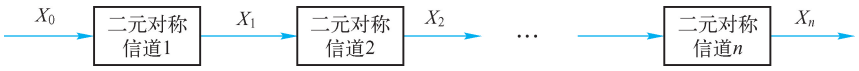
题图 3.5

（1）求最佳输入分布。

（2）求 $\varepsilon=\frac{1}{2}$时的信道容量。

（3）求 $\varepsilon \to 0$ 和 $\varepsilon \to 1$ 时的最佳输入分布值。

3.6 如题图 3.6 所示，把 n 个二元对称信道串接起来，每个二元对称信道的错误传递概率为 p。证明：这 n 个串接信道可以等效于一个二元对称信道，其错误传递概率为 $\frac{1}{2}[1-(1-2p)^n]$，且 $\lim_{n\to\infty} I(X_0;X_n)=0$（设 $p\neq 0$ 或 1）。



题图 3.6

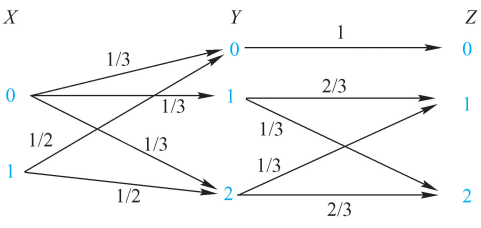
3.7 试求出准对称信道的信道容量的一般表达式。

3.8 试画出三元对称信道在理想（无噪声）和强噪声（输出不依赖于输入）情况下的信道模型，设信道输入等概分布。

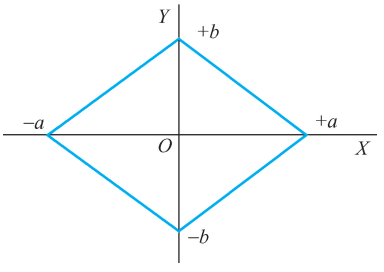
3.9 串联信道如题图 3.9 所示，求总信道矩阵。

3.10 设一时间离散、幅度连续的无记忆信道的输入是一个均值为零、方差为 E 的高斯随机变量，信道噪声为加性高斯噪声，方差为 $\sigma^2=1\ \mu W$，信道的符号传输速率为 $r=8000$ 符号/秒。如令一路电话通过该信道，电话机产生的信息率为 64 kbps，求输入信号功率 E 的最小值。

3.11 连续随机变量 X 和 Y 的联合概率密度函数在由 $\frac{1}{a}|x|+\frac{1}{b}|y|\leqslant 1$ 确定的菱形内均匀分布（如题图 3.11 所示）。



题图 3.9

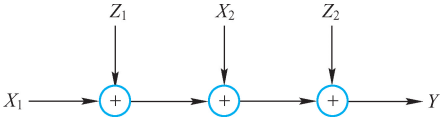


题图 3.11

（1）求 $I(X;Y)$。

（2）解释为什么 $I(X;Y)$ 与 a 和 b 无关。

3.12 一个高斯加性信道，其输入信号为 X_1 和 X_2，噪声信号为 Z_1 和 Z_2，输出信号为 $Y=X_1+X_2+Z_1+Z_2$，如题图 3.12 所示。输入和噪声均为相互独立的零均值的高斯随机变量，功率分别为 P_1、P_2 和 N_1、N_2。



题图 3.12

（1）求 $I(X_1;Y)$ 和 $I(X_2;Y)$。

（2）求 $I(X_1X_2;Y)$。

（3）当输入信号的总功率受限于 $P_1+P_2\leqslant P$ 时，求 $I(X_1;Y)+I(X_2;Y)$ 的最大值。

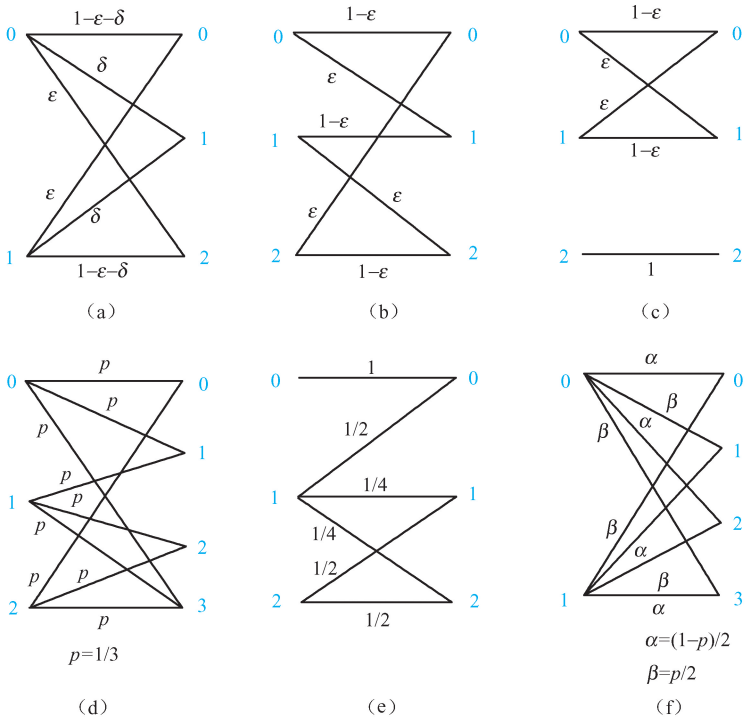
3.13 一个无记忆信道的输入为离散随机变量 X，噪声 Z 在区间 $[-a,+a]$ 上均匀分布，因此输出 $Y=X+Z$ 是一个连续随机变量。

（1）当 $X\in\{-1,1\}$ 且等概分布时，求 $I(X;Y)$（用 a 表示）。

（2）当 $X\in\{-1,0,1\}$ 且 $a=\frac{1}{2}$ 时，求最佳输入分布。

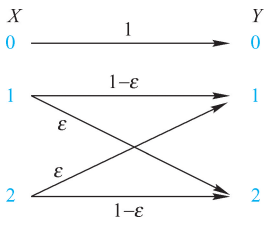
3.14 设某一信号的信息输出率为 5.6 kbps，噪声功率谱为 $N=5\times10^{-6}$ mW/Hz，在带宽 $B=4$ kHz 的高斯信道中传输。试求无差错传输需要的最小输入功率 P 是多少。

3.15 判断题图 3.15 中各信道是否对称；如对称，求出其信道容量。



题图 3.15

3.16 求题图 3.16 中信道的信道容量及其最佳的输入概率分布，并求当 $\varepsilon=0$ 和 $\frac{1}{2}$ 时的信道容量值。



题图 3.16

3.17 积信道。假设有两个离散无记忆信道 $\{X_1,p(y_1|x_1),Y_1\}$ 和 $\{X_2,p(y_2|x_2),Y_2\}$，信道容量分别为 C_1 和 C_2。两个信道同时输入 $x_1\in X_1$ 和 $x_2\in X_2$，输出 $y_1\in Y_1$ 和 $y_2\in Y_2$，这两个信道组成一个新的信道 $\{X_1\times X_2,p(y_1|x_1)\times p(y_2|x_2),Y_1\times Y_2\}$。求它的信道容量。

3.18 和信道。假设有两个离散无记忆信道 $\{X_1,p(y_1|x_1),Y_1\}$ 和 $\{X_2,p(y_2|x_2),Y_2\}$，信道

容量分别为 C_1 和 C_2。这两个信道的输入/输出符号集各不相同，并且假定每次只有一个信道有输入。

$$X\begin{cases} X_1 \to \boxed{p_1(y_1 \mid x_1)} \to Y_1 \\ X_2 \to \boxed{p_2(y_2 \mid x_2)} \to Y_2 \end{cases} Y$$

证明：

（1）$2^C = 2^{C_1} + 2^{C_2}$。

（2）如果 $C_1 > C_2$，那么一直用信道 1 是不是效率更高？

3.19 有记忆信道的信道容量高于无记忆信道的信道容量。考虑二元对称信道 $Y_i = X_i \oplus Z_i$，其中$\oplus$表示模 2 加，$X_i, Y_i \in \{0,1\}$。假定 $Z_1, Z_2, \cdots, Z_n$ 有相同的边缘概率分布 $\Pr\{Z_i = 1\} = p = 1 - \Pr\{Z_i = 0\}$，但并不相互独立，$Z^n$ 与输入 X^n 相互独立。

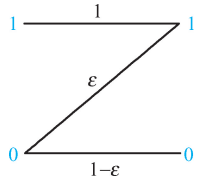
如果记 $C = 1 - H(p, 1-p)$，证明

$$\max_{p(x_1 x_2 \cdots x_n)} I(X_1 X_2 \cdots X_n ; Y_1 Y_2 \cdots Y_n) \geqslant nC$$

3.20 时变信道。离散无记忆时变信道 $p(y_1 \cdots y_n \mid x_1 \cdots x_n) = \prod_{i=1}^{n} p_i(y_i \mid x_i)$，求它的信道容量 $\max_{p(x_1 x_2 \cdots x_n)} I(X_1 X_2 \cdots X_n ; Y_1 Y_2 \cdots Y_n)$。

3.21 假定 C 为有 N 个输入、M 个输出的离散无记忆信道的信道容量，证明 $C \leqslant \min\{\log M, \log N\}$。

3.22 证明 n 个 $P_r\{Y=1 \mid X=0\} = \varepsilon$ 的 Z 信道（见题图 3.22）级联相当于一个 $P_r\{Y=1 \mid X=0\} = 1 - (1-\varepsilon)^n$ 的 Z 信道。



题图 3.22

3.23 一个信道的信道矩阵为

$$\begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} \frac{3}{4} & \frac{1}{4} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{1}{4} & \frac{3}{4} \end{bmatrix} \end{array}$$

（1）求信道容量。

（2）若它的输入概率分布为

$$P_r\{X=0\} = \frac{1}{2} - p$$

$$P_r\{X=1\} = 2p$$

$$P_r\{X=2\} = \frac{1}{2} - p$$

画出 $I(X=0;Y)$，$I(X=1;Y)$ 和 $I(X=2;Y)$ 作为 p 的函数的函数曲线，$0 \leqslant p \leqslant \frac{1}{2}$。你能从这些曲线得到关于这个信道的信道容量和最佳输入分布的什么结论？

3.24 假定 (X, Y, Z) 是一个多维高斯分布，且 $X \to Y \to Z$ 组成一个马尔可夫链，X 和 Y、Y 和 Z 的相关系数分别为 ρ_1 和 ρ_2，求 $I(X;Z)$。

第 4 章　无失真信源编码

对于信源来说有两个重要问题：一是信源输出的信息量的定量度量的问题，二是如何更有效地表示信源输出的问题。第 2 章回答了第一个问题，本章讨论第二个问题。信源输出的符号序列，经过信源编码，转换成适合信道传输的符号序列（一般称为码符号序列，如二元信道就要被编码为二元码序列），同时在不失真或允许一定失真的条件下，用尽可能少的码符号来传递信源消息，提高信息传输的效率。所以，信源编码是以提高通信的有效性为目的的，包括无失真信源编码和限失真信源编码。

本章主要讨论无失真信源编码，也就是在不允许失真的情况下怎样通过压缩信源的冗余度来提高信息传输率。无失真信源编码主要采用统计匹配编码，即根据信源符号的概率不同，编码的码长不同，概率大的信源符号的编码字短，概率小的信源符号的编码字长，这样可以使平均码长最短，达到压缩信源冗余度的目的。本章主要讲述的香农码、霍夫曼码、费诺码都是统计匹配编码。

4.1　信源编码概述

4.1.1　编码器

将信源符号序列按一定的数学规律映射成码符号序列的过程称为信源编码，完成编码功能的器件称为编码器。接收端有一个译码器完成相反的功能。图 4.1 所示为信源编码模型。

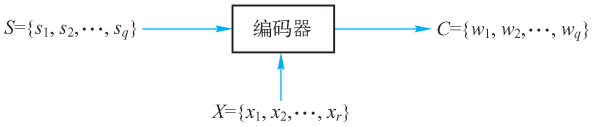


图 4.1　信源编码模型

信源编码器的输入是信源符号集 $S=\{s_1,s_2,\cdots,s_q\}$，共 q 个信源符号。同时存在另一个符号集，称为码符号集 $X=\{x_1,x_2,\cdots,x_r\}$，共 r 个码符号，码符号集中的元素称为码元或码符号，编码器的作用就是将信源符号集 S 中的符号 $s_i(i=1,2,\cdots,q)$ 变换成由 l_i 个码符号组成的一一对应的输出符号序列 $w_i=x_{i_1}x_{i_2}\cdots x_{i_{l_i}}$。编码器输出的符号序列又称为码字，并用 w_i $(i=1,2,\cdots,q)$ 来表示，它与信源符号 $s_i(i=1,2,\cdots,q)$ 是一一对应的。

码字的集合 C 称为码，即 $C=\{w_1,w_2,\cdots,w_q\}$。信源符号 s_i 对应码字 w_i，要用 l_i 个码符号来表示。l_i 称为码字长度，简称码长。

所以，信源编码就是把信源符号序列变换到码符号序列的一种映射。若要实现无失真编

码，那么这种映射必须是一一对应的，是可逆的。一般来说，人们总是希望把信源输出的信息毫无保留地传输到接收端，即实现无失真传输，所以要对信源实现无失真编码。

无失真信源编码主要针对离散信源，连续信源在量化编码的过程中必然会有量化失真，所以对连续信源只能近似地再现信源的消息。

由第 2 章的讨论可知，由于信源概率分布的不均匀性和符号之间存在的相关性，使得信源存在冗余度，而实际上传输信源信息只需要传输信源极限熵大小的信息量。信源编码的主要任务是减少冗余度，提高编码效率，具体地说，就是针对信源输出符号序列的统计特性，寻找一定的方法，把信源输出符号序列变换为最短的码字序列。去除冗余度的方法有两个：一是去除相关性，使编码后码序列的每个码符号尽可能互相独立，一般是利用对信源的符号序列进行编码，而不是对单个的信源符号进行编码的方法实现；二是使编码后每个码符号出现的概率尽可能相等，即概率分布均匀化，可以通过概率匹配的方法来实现，也就是使小概率消息对应长码，大概率消息对应短码。

下面举一个例子说明怎样用编码器实现对信源符号的编码。

【例 4.1】信源概率空间为

$$\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & \cdots & s_4 \\ p(s_1) & p(s_2) & \cdots & p(s_4) \end{bmatrix}$$

把信源符号编码为适合二元信道的二元码。二元信道是数字通信中最常用的一种信道，它的码符号集为$\{0,1\}$。

【解】

将信源通过一个二元信道传输，就必须把信源符号 s_i 变换成由 0、1 符号组成的码符号序列，即进行编码。我们可用不同的二元码符号序列 w_i 与信源符号 s_i 一一对应，这样就可以得到不同的码。一般情况下，码可分为两类：固定长度码和可变长度码。固定长度码又称为定长码，码组中所有码字的长度都相同，如表 4.1 中的码 1。可变长度码又称为变长码，码组中所有码字的长短不一，即码字中的码符号的个数不同，如表 4.1 中的码 2。

本章讨论的都是同价码，即每个码符号所占的传输时间都相同的码。显然，对同价码来说，定长码中的每个码字 $w_i(i=1,2,\cdots,q)$ 的传输时间相同，而变长码中每个码字的传输时间不一定相等。

一般，信源输出的是符号序列，把离散无记忆信源的 N 次扩展信源的概念加以引伸，便得到 N 次扩展码。假定把信源 $S=\{s_1,s_2,\cdots,s_q\}$ 中的符号变换成码 $C=\{w_1,w_2,\cdots,w_q\}$ 中的码字，则 N 次扩展信源 $\boldsymbol{S}=\{\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_{q^N}\}$ 对应 N 次扩展码 $\boldsymbol{C}=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_{q^N}\}$。例如，表 4.1 中的码 2 的二次扩展码如表 4.2 所示。

表 4.1　二元码

信源符号 s_i	$p(s_i)$	码 1	码 2
s_1	$p(s_1)$	00	0
s_2	$p(s_2)$	01	01
s_3	$p(s_3)$	10	001
s_4	$p(s_4)$	11	111

表 4.2　二次扩展码

二次扩展信源符号 s_j	二次扩展码字 w_j
$\boldsymbol{s}_1 = s_1\,s_1$	00
$\boldsymbol{s}_2 = s_1\,s_2$	001
$\boldsymbol{s}_3 = s_1\,s_3$	0001
…	…
$\boldsymbol{s}_{16} = s_4\,s_4$	111111

4.1.2 码的分类

1. 分组码和非分组码

如表 4.2 的二次扩展码所示，信源编码过程可以抽象为一种映射，即将信源符号集 $S=\{s_i\}(i=1,2,\cdots,q)$ 中的每个信源符号 s_i 分别映射为一个长度为 l_i 的码字 $w_i(i=1,2,\cdots,q)$，而信源符号序列对应的码符号序列是由这些信源符号对应的码字组合而成的。

【定义 4-1】将信源符号集中的每个信源符号 s_i 固定地映射成一个码字 w_i，这样的码称为分组码，又称为块码。

直观地理解，分组码就是把信源的符号序列分成组，从信源符号序列到码字序列的变换是在分组的基础上进行的，特定的信源符号组唯一地确定了特定的码字组。

与分组码相对应的分类是非分组码，又称为树码。树码编码器输出的码符号通常与编码器的所有输入的信源符号都有关，如后面会讲到的算术编码。

2. 奇异码和非奇异码

【定义 4-2】若一种分组码中的所有码字都不相同，则称此分组码为非奇异码， 否则称为奇异码。

在表 4.3 中，码 1 是奇异码，码 2 是非奇异码。但分组码为非奇异码是正确译码的必要条件，而不是充分条件。例如传送分组码 2 时，如果接收端接收到 00 时，并不能确定发送端的消息是 s_1s_2 还是 s_3。

表 4.3　奇异码和非奇异码

信源符号 s_i	码 1	码 2
s_1	0	0
s_2	11	10
s_3	00	00
s_4	11	01

3. 唯一可译码和非唯一可译码

【定义 4-3】任意有限长的码元序列，只能够唯一地被分割成一个个码字，便称为唯一可译码。

唯一可译码的含义是指，不但要求不同的码字表示不同的信源符号，而且进一步要求对由信源符号构成的符号序列进行编码时，在接收端仍能正确译码而不发生混淆。唯一可译码首先是非奇异码，且任意有限长的码字序列不会雷同。一个分组码若对于任意有限的整数 N，其 N 次扩展码均为非奇异的，则为唯一可译码。

4. 即时码和非即时码

同是唯一可译码，其译码方法仍有不同。如表 4.4 中列出的两组唯一可译码，其译码方法不同。当传输码 1 时，信道输出端接收到一个码字后不能立即译码，必须等到下一个码字接收到时才能判断是否可以译码。若传输码 2，则无此限制，接收到一个完整码字后立即可译码。后一种码被称为即时码。

表 4.4　即时码和非即时码

信源符号 s_i	码 1	码 2
s_1	1	1
s_2	10	01
s_3	100	001
s_4	1000	0001

【定义 4-4】不需考虑后续的码符号就可以从码符号序列中译出码字，这样的唯一可译码称为即时码。

下面讨论唯一可译码成为即时码的条件。

【定义 4-5】设 $w_i = x_{i_1}x_{i_2}\cdots x_{i_l}$ 为一码字，对于任意 $1 \leqslant j \leqslant l$，称码符号序列的前 j 个元素 $x_{i_1}x_{i_2}\cdots x_{i_j}$ 为码字 w_i 的前缀。

按照上述的前缀的定义，有下述结论。

【定理 4-1】一个唯一可译码成为即时码的充要条件是，其中任何一个码字都不是其他码字的前缀。

【证明】

这个很好理解，因为如果没有一个码字是其他码字的前缀，则在接收到一个相当于一个完整码字的码符号序列后便可以立即译码，而无须考虑其后的码符号。反过来说，如果有一个码字是其他码字的前缀，假设 w_j 是 w_i 的前缀，则在收到相当于 w_j 的码符号序列后，还不能立即判定它是一个完整的码字，若想正确译码，还必须参考后续的码符号，这与即时码的定义相矛盾，所以即时码的必要条件是其中任何一个码字都不是其他码字的前缀。

综上所述，可将码进行如下分类：

码
- 非分组码
- 分组码
  - 奇异码
  - 非奇异码
    - 非唯一可译码
    - 唯一可译码
      - 即时码
      - 非即时码

5. 树图

即时码可以方便地用树图来构造。图 4.2 所示是一个二元即时码。

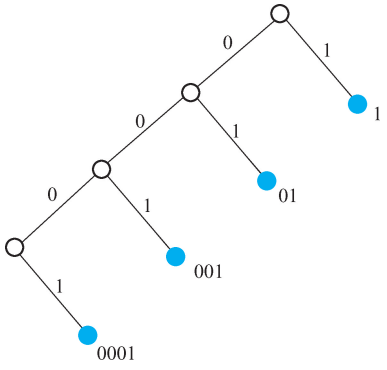


图 4.2　二元即时码的树图

树图中有树枝和节点。树图最顶部的节点称为根节点，不再产生分支的节点称为终端节点，其他节点称为中间节点。每个节点能够生出的树枝数目等于码符号数 r（这里 $r=2$）。在树枝旁边分别标以相应的码符号 $0,1,\cdots,r-1$。将从根节点到终端节点的各树枝代表的码符号顺次连接，就得到了编码码字。

树图中自根部经过一个分枝到达的节点称为一阶节点，一阶节点最多有 r 个。它们的子

节点称为二阶节点，二阶节点最多有 r^2 个。以此类推，N 阶节点最多有 r^N 个。

若指定某个 N 阶节点为终端节点，表示一个信源符号，则该节点就不再延伸，相应的码字即为从树根到此端点的树枝标号序列，其长度为 N。这样构造的码满足即时码的条件，因为从树根到每个终端节点所走的路径均不相同，并且中间不安排码字，故一定满足对前缀的限制。如果有 q 个信源符号，那么在码树上就要选择 q 个终端节点，相应地有 q 个码字。

即时码的码树图还可以用来译码。当接收到一串码符号序列后，首先从树根出发，根据接收到的第一个码符号来选择应走的第一条路径，沿着所选支路如果走到一个中间节点，再根据接收到的第二个码符号来选择应走的第二条路径，若又走到中间节点，就依次继续下去，直到终端节点为止。走到终端节点，就可根据所走的支路立即判断出所接收的码字，同时使系统重新返回树根，再做下一个接收码字的判断。这样可以将接收到的一串码符号序列译成对应的信源符号序列。

4.2 定长码及定长信源编码定理

一般来说，若要实现无失真的编码，所编的码必须是唯一可译码，否则会因译码带来错误与失真。对于定长码来说，若定长码是非奇异码，则它的任意有限长 N 次扩展码一定也是非奇异码，因此定长非奇异码一定是唯一可译码。表 4.5 中的码 2 是奇异码，当接收到码符号 11 后，既可译成 s_2 也可译成 s_4，所以不能唯一地译码。码 1 是等长非奇异码，因此它是一个唯一可译码。

表 4.5 定长码

信源符号 s_i	码 1	码 2
s_1	00	00
s_2	01	11
s_3	10	10
s_4	11	11

若对一个有 q 个信源符号的信源 S 进行定长编码，那么信源 S 存在唯一可译定长码的条件是 $q \leqslant r^l$，其中 r 是码符号集中的码元数，l 是定长码的码长。例如在表 4.5 中，信源 S 共有 $q=4$ 个信源符号，进行二元等长编码，$r=2$，则信源存在唯一可译定长码的条件是 $l \geqslant 2$。

如果对信源 S 的 N 次扩展信源进行定长编码，若要编得的定长码是唯一可译码，则必须满足

$$q^N \leqslant r^l \tag{4.1}$$

其中，q^N 是信源 S 的 N 次扩展信源的符号个数。当把 N 次扩展信源的信源符号 $\boldsymbol{s}_j$ $(j=1,2,\cdots,q^N)$ 编成长度为 l 的码字时，如果要求编得的定长码是唯一可译码，则每个信源符号 $\boldsymbol{s}_j$ 都有一个码字对应，所以必须满足式(4.1)。换句话说，只有当 l 长的码符号序列个数 r^l 不小于 N 次扩展信源的符号个数 q^N 时，才能存在定长非奇异码。

对式(4.1)两边取对数有

$$N\log q \leqslant l\log r$$

得

$$\frac{l}{N} \geqslant \frac{\log q}{\log r} = \log_r q$$

$\frac{l}{N}$表示平均每个原始信源符号所需要的码符号个数，对于定长唯一可译码，平均每个原始信源符号至少需要用$\log_r q$ 个码符号来表示。

当 $r=2$ 时，$\frac{l}{N} \geqslant \log q$，表示对于二元定长唯一可译码，平均每个原始信源符号至少需要用 $\log q$ 个二元符号来表示。当 $N=1$ 时，则有 $l \geqslant \log q$。

例如，英文电报信源有 32 个符号（26 个英文字母加 6 个标点符号），即 $q=32$，对每个符号进行二元编码，即 $r=2$，则 $l \geqslant \log_r q = \log_2 q = \log_2 32 = 5$。也就是说，每个英文电报符号至少要用 5 位二元符号进行编码才能得到唯一可译码。

由第 2 章已知，实际英文电报信源，在考虑了符号出现的概率以及符号之间的相关性以后，平均每个英文电报符号所提供的信息量约等于 1.4 比特，远小于 5 比特。因此，定长编码后，每个码字只载荷约 1.4 比特信息量，也就是 5 个二元码符号只携带约 1.4 比特信息量，而 5 个二元码符号最大能载荷 5 比特的信息量。因此，定长编码的信息传输效率是很低的。那么怎么才能提高信息传输效率，也就是怎样才能使每个信源符号用尽可能少的码符号来表示呢？

方法 1：对信源符号序列进行编码，在考虑符号间的依赖关系后，有些信源符号序列不会出现，这样可能出现的信源符号序列个数会小于 q^N。对于不会出现的符号序列不予编码，这样不会造成误差。

方法 2：对于概率非常小的信源符号序列不予编码，这样可能造成一定的误差，但是当信源符号序列长度 N 足够大时，这种误差概率可以任意小，即可做到几乎无失真编码。

下面将讨论的定长信源编码定理给出了定长信源编码所需码长的理论极限值。

定长信源编码定理的证明中，需要用到渐进等分割性和 ε 典型序列的概念，具体请参见附录 A。

离散无记忆信源 S 的 N 次扩展信源的输出序列 $\boldsymbol{s}_j$ 可以分为两类，一类是 ε 典型序列，另一类是非 ε 典型序列。ε 典型序列是 $\frac{I(\boldsymbol{s}_j)}{N}$ 非常接近 $H(S)$ 的一类序列。

可以证明，当 $N \to \infty$ 时，e 典型序列的概率总和趋于 1，因此 ε 典型序列的总概率大，非 ε 典型序列的总概率小（注意，单个非 ε 典型序列出现的概率不一定比 ε 典型序列出现的概率小）。

虽然 ε 典型序列的总概率很大，但是它的个数在全部序列中的比例却很小，因此我们只对少数的 ε 典型序列进行等长编码，此时码字总数会大大减少，所需码长也得以减少。

由 ε 典型序列的性质，我们可以很容易地推出以下定长信源编码定理。

【定理 4-2】离散无记忆信源的熵为 $H(S)$，若对信源长为 N 的序列进行定长编码，码符号集中有 r 个码符号，码长为 l，则对于任意 $\varepsilon > 0$，只要满足

$$\frac{l}{N} \geqslant \frac{H(S)+\varepsilon}{\log r}$$

则当 N 足够大时，可实现几乎无失真编码，即译码错误概率为任意小。反之，如果

$$\frac{l}{N} \leqslant \frac{H(S)-2\varepsilon}{\log r}$$

则不可能实现几乎无失真编码，且当 N 足够大时，译码错误概率为 1。

定长信源编码定理给出了定长编码时每个信源符号最少所需的码符号的理论极限，该极限值由 $H(S)$ 决定。当二元编码时 $r=2$，有

$$\frac{l}{N} \geqslant H(S)+\varepsilon \tag{4.2}$$

这是定长编码时平均每个信源符号所需的二元码符号数的理论极限。

定理 4-2 是在离散平稳无记忆信源的条件下证明的，但它同样适合于平稳有记忆信源，只要将式中的 $H(S)$ 改为极限熵 H_∞ 即可，条件是有记忆信源的极限熵 H_∞ 和极限方差存在。

比较式(4.2)与 $\frac{l}{N}=\log q$，当信源符号等概分布时，两式就完全一致，但一般情况下信源符号并非等概分布，而且符号之间有很强的关联性，故信源的熵 H_∞（极限熵）远小于 $\log q$。根据定长信源编码定理，每个信源符号平均所需的二元码符号可大大减少，从而使编码效率提高。

仍以英文电报为例，由于英文信源的极限熵 $H_\infty \approx 1.4$ 比特/信源符号，所以码长 l 只要满足 $\frac{l}{N}>1.4$，即平均每个英文符号只需近似用 1.4 个二元码符号来表示，这比由 $\frac{l}{N}\geqslant \log q$ 计算的需要 5 个二元符号减少了许多，从而提高了信息传输效率。

定理 4-2 的条件又可写成

$$l\log r > NH(S)$$

该式表明只要 l 长码符号序列所能携带的最大信息量大于 N 长信源序列所携带的信息量，就可以实现无失真编码，当然条件是 N 足够大。

为了衡量编码效果，我们引进编码效率的概念。

【定义 4-6】设熵为 $H(S)$ 的离散无记忆信源，若对信源的长为 N 的符号序列进行定长编码，码符号集中的码符号个数为 r，设码字长为 l，定义 $\eta=\frac{H(S)}{\frac{l}{N}\log r}$ 为编码效率。其中，$\frac{l}{N}\log r$（比特/信源符号）表示平均每个信源符号编码后能载荷的最大信息量。

因此，定理 4-2 的条件也可表述为，平均每个信源符号编码后能载荷的最大信息量大于信源的熵才能实现几乎无失真编码。

根据定理 4-2，最佳编码效率可表示为

$$\eta=\frac{H(S)}{H(S)+\varepsilon}\quad \varepsilon>0$$

所以

$$\varepsilon=\frac{1-\eta}{\eta}\times H(S)$$

在已知方差 $D[I(s_i)]$ 和信源熵的条件下，要达到最佳编码效率 η 且允许的译码错误概率要小于任一给定的正数 δ，信源序列长度 N 须满足

$$N\geqslant\frac{D[I(s_i)]}{\varepsilon^2\delta}=\frac{D[I(s_i)]}{H^2(S)}\times\frac{\eta^2}{(1-\eta^2)\delta} \tag{4.3}$$

从式(4.3)可以看出，既要允许错误概率 δ 小，又要编码效率高，那么信源序列长度 N 必须越长。在实际情况下，要实现几乎无失真的定长编码，N 需要的长度将会大到难以实现。

【例 4.2】设有离散无记忆信源

$$\begin{bmatrix}S\\P\end{bmatrix}=\begin{bmatrix}s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 & s_8\\0.4 & 0.18 & 0.10 & 0.10 & 0.07 & 0.06 & 0.05 & 0.04\end{bmatrix}$$

如果对信源符号采用定长二元编码，要求编码效率 $\eta=90\%$，允许错误概率 $\delta\leqslant 10^{-6}$，求所需信源序列的长度 N。

【解】

信息熵

$$\begin{aligned}H(S)&=E[-\log p(s_i)]\\&=-\sum_{i=1}^{8}p(s_i)\log p(s_i)\\&=2.55\text{ 比特/信源符号}\end{aligned}$$

自信息的方差

$$D[I(s_i)]=\sum_{i=1}^{8}p(s_i)[-\log p(s_i)]^2-H(S)^2=7.82$$

$$\varepsilon=\frac{1-\eta}{\eta}H(S)=0.28$$

所以

$$\begin{aligned}N&\geqslant\frac{D[I(s_i)]}{\varepsilon^2\delta}=\frac{7.82}{0.28^2\times 10^{-6}}\\&=9.8\times 10^7\approx 10^8\end{aligned}$$

即信源序列长度 N 需长达 10^8 以上才能实现上述给定的要求。

【例 4.3】设离散无记忆信源 $\begin{bmatrix}S\\P\end{bmatrix}=\begin{bmatrix}s_1 & s_2\\\frac{3}{4} & \frac{1}{4}\end{bmatrix}$，要求 $\eta=0.96$，$\delta\leqslant 10^{-5}$，求 N。

【解】信源熵

$$H(S)=\frac{1}{4}\log 4+\frac{3}{4}\log\frac{4}{3}=0.811\text{ 比特/信源符号}$$

自信息的方差

$$\begin{aligned}D[I(s_i)]&=\sum_{i=1}^{2}p(s_i)[-\log p(s_i)]^2-H(S)^2\\&=\frac{3}{4}\left(\log\frac{3}{4}\right)^2+\frac{1}{4}\left(\log\frac{1}{4}\right)^2-(0.811)^2\\&=0.4715\end{aligned}$$

因为 $\varepsilon=\frac{1-\eta}{\eta}H(S)$，所以

$$N\geqslant\frac{0.4715}{(0.811)^2}\frac{(0.96)^2}{(0.04)^2\times 10^{-5}}=4.13\times 10^7$$

即信源序列长度长达 4×10^7 以上，才能实现给定的要求，这在实际中是很难实现的。因此，一般来说，当 N 有限时，高编码效率的定长码往往要引入一定的失真和错误，但是变长码则可以在 N 不大时就可以实现无失真编码。

4.3 变长码及变长信源编码定理

变长码要成为唯一可译码，不但要求其本身应是非奇异的，而且其有限长 N 次扩展码

也应是非奇异的。例如，表 4.6 中的码 2 是非奇异码，但是当信宿收到“00”时，它不能判断是 s_1s_1 还是 s_3，所以不是唯一可译码。码 3、码 4 是唯一可译码，其中码 4 还是即时码。什么样的码是唯一可译码或即时码呢？下面的 Kraft 不等式和 McMillan 不等式分别给出了即时码和唯一可译码的码长条件。

表 4.6　变长码

信源符号 s_i	概率分布	码 1	码 2	码 3	码 4
s_1	$\frac{1}{2}$	0	0	1	1
s_2	$\frac{1}{4}$	11	10	10	01
s_3	$\frac{1}{8}$	00	00	100	001
s_4	$\frac{1}{8}$	11	01	1000	0001

4.3.1　Kraft 不等式和 McMillan 不等式

对于一个给定信源的编码，Kraft 不等式和 McMillan 不等式可以让我们从码长来判断它是否满足即时码和唯一可译码的条件。这两个不等式在形式上是完全一样的。

【定理 4-3】设信源符号集为 $S=\{s_1,s_2,\cdots,s_q\}$，码符号集为 $X=\{x_1,x_2,\cdots,x_r\}$，对信源进行编码，得到的码为 $C=\{w_1,w_2,\cdots,w_q\}$，码长分别为 $l_1,l_2,\cdots,l_q$，则即时码存在的充要条件是

$$\sum_{i=1}^{q} r^{-l_i} \leqslant 1 \tag{4.4}$$

这称为 Kraft 不等式。

【证明】

（1）充分性

证明如果码长满足不等式(4.4)，则可以得到即时码。

现在假设码字长度满足 $l_1 \leqslant l_2 \leqslant \cdots \leqslant l_q \leqslant l$，由于这样的假设只影响码字的编号，所以是可以的。码长最长为 l，深度为 l 的 r 叉树，如果所有分枝都延伸到最后一节则有 r^l 个终端节点，可以编 r^l 个码字。如果从根出发，取一个 l_1 阶节点作为码字 w_1，则后续的树枝应该被砍去，因此共有 r^{l-l_1} 个 l 阶节点不能得到使用。

同样，在 l_i 阶节点上的终端节点可以使得 r^{l-l_1} 个 l 阶节点不能得到使用。最后 l 阶节点只剩下 $r^l - r^{l-l_1} - r^{l-l_2} \cdots - r^{l-l_q} = r^l - \sum_{i=1}^{q} r^{l-l_i}$ 个可以使用的终端节点。

如果码长满足不等式(4.4)，即 $1 - \sum_{i=1}^{q} r^{-l_i} \geqslant 0$，那么 $r^l - \sum_{i=1}^{q} r^{l-l_i} \geqslant 0$，即可以使用的 r 阶节点数大于等于 0，可以构造一个即时码。

（2）必要性

证明即时码必然满足不等式(4.4)。由于即时码必然可以用树图来构造，并且安排了码字的终端节点不会再生出树枝，我们可取一个有 l 阶节点的 r 叉树且 $l \geqslant \max_i l_i$。树的第 0 节为根，在第 l 阶上可有 r^l 个节点。长为 l_i 的码字相当于砍去了第 l 阶上的 r^{l-l_i} 个节点，q 个码字共砍去第 l 阶的节点数必小于 r^l，即

$$\sum_{i=1}^{q} r^{l-l_i} \leqslant r^l$$

那么

$$\sum_{i=1}^{q} r^{-l_i} \leqslant 1$$

证毕。

由 Kraft 不等式可知，给定 r、q，只要允许码字长度可以足够长，则码长总可以满足 Kraft 不等式而得到即时码，Kraft 不等式指出了即时码的码长必须满足的条件。后来，McMillan 证明唯一可译码的码长也必须满足此不等式。在码长的选择上唯一可译码并不比即时码有更宽松的条件。因此，对于唯一可译码，该不等式又称为 McMillan 不等式。

【定理 4-4】 唯一可译码存在的充要条件是

$$\sum_{i=1}^{q} r^{-l_i} \leqslant 1$$

其中，r 为码符号个数，l_i 为码字长度，q 为信源符号个数。这称为 McMillan 不等式。

定理 4-4 指出了唯一可译码中 r、q、l_i 之间的关系，如果码长满足这个不等式，则一定能够构造至少一种唯一可译码，否则，无法构成唯一可译码。

例如在表 4.6 中，码 1、码 2 的码长为 $l_1=1$，$l_2=l_3=l_4=2$，$\sum_{i=1}^{4} 2^{-l_i}=\frac{5}{4}>1$，所以不能构成唯一可译码。而码 3、码 4 的 $l_1=1$，$l_2=2$，$l_3=3$，$l_4=4$，$\sum_{i=1}^{4} 2^{-l_i}=\frac{15}{16}<1$，可以构成唯一可译码。当然，满足此码长条件的也不一定就是唯一可译码，如 $C=\{1,01,011,0001\}$，但至少可以找到一个是唯一可译码。同样，满足此码长条件的码也可能不是即时码，但至少可以找到一种即时码。

从定理 4-3 和定理 4-4 可以得到一个重要的结论：任何一个唯一可译码均可用一个相同码长的即时码来代替。因为即时码很容易用树图法构造，所以要构造唯一可译码，只要构造即时码即可。用树图法可以方便地实现即时码的编码和译码，也可以方便地判别一个码是不是即时码，那么唯一可译码怎么判别呢？

4.3.2 唯一可译码的判别准则

定理 4-4 只给出了唯一可译码的码长条件，不能作为一个具体的码是否为唯一可译码的判别方法，因为满足 McMillan 不等式的码不一定是唯一可译码，也不可能根据唯一可译码的定义来判断，因为在实际应用中不可能一一检查所有 N 阶扩展码的奇异性。A. A. Sardinas 和 G. W. Patterson 于 1957 年提出了一种唯一可译码的判别准则，其原理如图 4.3 所示。其中 A_i、B_i 都是码字，如果某个有限长的码符号序列能译成两种不同的码字序列，此码不是唯一可译码，此时 B_1 一定是 A_1 的前缀，而 A_1 的尾随后缀一定是另一码 B_2 的前

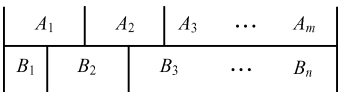


图 4.3 唯一可译码的判别准则

缀；而 B_2 的尾随后缀又是其他码字的前缀。最后，码符号序列的尾部也一定是一个码字。

根据这个原理，该判别准则的判别方法如下。

设 C 为码字的集合，我们要构造尾随后缀的集合 $F_1, F_2, \cdots, F$。

① F_1 是 C 中所有码字尾随后缀的集合：若 C 中码字 w_j 是码字 w_i 的前缀，即 $w_i = w_j A$，则将尾随后缀 A 列为 F_1 中的元素，所有这样的尾随后缀构成了 F_1。

② 考察 C 和 F_i 两个集合，如果 C 中任一码字是 F_i 中元素的前缀，或者 F_i 中任一元素是 C 中码字的前缀，则将其相应的尾随后缀放入集合 F_{i+1}。

③ $F = \bigcup_i F_i$。

④ 一旦 F 中出现了 C 中的元素，则算法终止，判断 C 不是唯一可译码；若出现 F_{i+1} 为空集，算法终止，判断 C 是唯一可译码。

即一个码是唯一可译码的充要条件是 F 中不含有 C 中的码字。

【例 4.4】消息集合 $\{s_1, s_2, s_3, s_4, s_5, s_6, s_7\}$ 有 7 个元素，分别被编码为 $\{a, c, ad, abb, bad, deb, bbcde\}$，判断是否是唯一可译码。

【解】

按照上述方法构造表 4.7 所列的尾随后缀集。

表 4.7　唯一可译码的判别准则

C	F_1	F_2	F_3	F_4	F_5
a	d	eb	de	b	ad
c	bb	cde			$bcde$
ad					
abb					
bad					
deb					

F_5 中的第一个元素正好是 C 中的第三个码字，所以 C 不是唯一可译码。

4.3.3　紧致码平均码长界限定理

由前面的讨论可知，对同一信源用同一码符号集编成的即时码或唯一可译码可有很多种，究竟哪一种最好呢？从高效传输信息的角度，我们希望尽可能选择码符号序列长度较短的码字，也就是采用码长较短的码字，为此我们引入码的平均长度的概念。

【定义 4-7】设有信源

$$\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & \cdots & s_q \\ p(s_1) & p(s_2) & \cdots & p(s_q) \end{bmatrix}$$

编码后的码字分别为 $w_1, w_2, \cdots, w_q$，各码字相应的码长分别为 $l_1, l_2, \cdots, l_q$。因为是唯一可译码，信源符号 s_i 和码字 w_i 一一对应，则定义此码的平均码长为

$$\bar{L} = \sum_{i=1}^{q} p(s_i) l_i \text{ 码符号/信源符号}$$

$\bar{L}$ 表示每个信源符号平均需用的码元数。

【定义 4-8】当信源给定时，信源熵 $H(S)$ 就确定了，而编码后每个信源符号平均用 $\bar{L}$ 个码元来表示，编码后平均每个码元载荷的信息量 $H(X)$ 定义为编码后的信息传输率 R：

$$R = H(X) = \frac{H(S)}{\bar{L}} \text{比特/码符号}$$

如果传输一个码符号平均需要 t 秒时间，则编码后信源每秒钟提供的信息量为

$$R_t = \frac{H(S)}{\bar{L} \times t} \text{比特/秒}$$

R_t 越大，信息传输率就越高，因此我们感兴趣的码是使平均码长 $\bar{L}$ 为最短的码。

【定义 4-9】对于给定的信源和码符号集，若有一个唯一可译码，其平均码长 $\bar{L}$ 小于所有其他唯一可译码的平均码长，则称这个码为紧致码或最佳码。

无失真信源编码的核心问题就是寻找紧致码。

下面来看紧致码的平均码长 $\bar{L}$ 可能达到的理论极限。

【定理 4-5】若一个离散无记忆信源 S，其熵为 $H(S)$，码符号集 $X = \{x_1, x_2, \cdots, x_r\}$，总可以找到一种唯一可译码，其平均码长满足

$$\frac{H(S)}{\log r} \leqslant \bar{L} < 1 + \frac{H(S)}{\log r} \tag{4.5}$$

该定理说明，平均码长 $\bar{L}$ 不能小于极限值 $\frac{H(S)}{\log r}$，否则唯一可译码不存在，同时定理又给出了平均码长的上界。这不是说大于该上界就不能构成唯一可译码，但是希望 $\bar{L}$ 尽可能短，因此我们关心的是紧致码的平均码长的范围。$\bar{L}$ 是紧致码的平均码长，它与 $H(S)$ 有关。

另外，这个极限值与定长信源编码定理中的极限值是一致的。下面证明这个定理。

【证明】

先证下界成立，即证 $H(S) - \bar{L}\log r \leqslant 0$。

$$\begin{aligned}
H(s) - \bar{L}\log r &= -\sum_{i=1}^{q} p(s_i)\log p(s_i) - \sum_{i=1}^{q} p(s_i) l_i \log r \\
&= -\sum_{i=1}^{q} p(s_i)\log p(s_i) + \sum_{i=1}^{q} p(s_i)\log r^{-l_i} \\
&= \sum_{i=1}^{q} p(s_i)\log \frac{r^{-l_i}}{p(s_i)} \\
&\leqslant \log \sum_{i=1}^{q} \left[p(s_i) \frac{r^{-l_i}}{p(s_i)} \right] \\
&= \log \sum_{i=1}^{q} r^{-l_i}
\end{aligned}$$

根据 McMillan 不等式，唯一可译码的码长满足 $\sum_{i=1}^{q} r^{-l_i} \leqslant 1$，所以 $\bar{L} \geqslant \frac{H(s)}{\log r}$ 成立。

等号成立的条件是对于 $\forall i$，$\frac{r^{-l_i}}{p(s_i)} = 1$，即 $p(s_i) = r^{-l_i}$，这时

$$H(S) - \bar{L}\log r = \sum_{i=1}^{q} p(s_i)\log 1 = 0$$

因此

$$\bar{L} = \frac{H(S)}{\log r}$$

这时每个码字的相应码长

$$l_i = \frac{-\log p(s_i)}{\log r} = -\log_r p(s_i)$$

下面证明上界成立，即证明满足条件 $\bar{L} < 1 + \frac{H(S)}{\log r}$ 的唯一可译码是存在的。

从上面的证明过程知道，平均码长要达到下界，必须满足

$$l_i = \frac{-\log p(s_i)}{\log r} = -\log_r p(s_i)$$

即每个信源符号的概率分布恰好使它的码长为整数，这是比较苛刻的条件，如果不能满足的话，那么 l_i 应该是小于 $-\log_r p(s_i)+1$ 的一个整数，即

$$l_i < -\log_r p(s_i) + 1$$

所以

$$p(s_i) < r^{-(l_i-1)} \qquad i = 1,2,\cdots,q$$

对这个不等式略加变化可得

$$\sum_{i=1}^{q} p(s_i)\log p(s_i) < \sum_{i=1}^{q} p(s_i)\log r^{-(l_i-1)}$$

$$H(S) > \log r \sum_{i=1}^{q} p(s_i)(l_i - 1) = \log r(\bar{L} - 1)$$

所以

$$\bar{L} < \frac{H(S)}{\log r} + 1$$

综合上、下界，得到

$$\frac{H(S)}{\log r} \leqslant \bar{L} < \frac{H(S)}{\log r} + 1$$

若熵以 r 进制为单位，则式(4.6)可写成

$$H_r(S) \leqslant \bar{L} < H_r(S) + 1$$

4.3.4 无失真变长信源编码定理(香农第一定理)

与无失真定长信源编码定理一样，这也是一个关于码长的极限定理，这个极限和定长编码的极限其实是一致的。

【定理 4-6】设有离散无记忆信源 S，信源熵为 $H(S)$，其 N 次扩展信源为 $S^N = \{s_1, s_2, \cdots, s_{q^N}\}$，熵为 $H(S^N)$，并用码符号 $X = \{x_1, x_2, \cdots, x_r\}$ 对信源 S^N 进行变长编码，那么总可以找到一种唯一可译码，使每个信源符号所需的平均码长 $\frac{\bar{L}_N}{N}$ 满足

$$\frac{H(S)}{\log r} \leqslant \frac{\bar{L}_N}{N} < \frac{H(S)}{\log r} + \frac{1}{N}$$

或

$$H_r(S) \leqslant \frac{\bar{L}_N}{N} < H_r(S) + \frac{1}{N}$$

当 $N \to \infty$ 时，则

$$\frac{\bar{L}_N}{N} = H_r(S)$$

其中，$\bar{L}_N = \sum_{i=1}^{q^N} p(s_i)\lambda_i$。其中 λ_i 是扩展信源的信源符号 $\boldsymbol{s}_i$ 所对应的码字长度，因此$\bar{L}_N$是扩展信源的信源符号的平均码长，而$\frac{\bar{L}_N}{N}$仍是单个信源符号所需的平均码长。

这里要注意$\frac{\bar{L}_N}{N}$与 $\bar{L}$ 的区别：它们都是单个信源符号所需的码符号的平均数，但$\frac{\bar{L}_N}{N}$的含义是，为了得到这个平均值，不是对单个信源符号 $\boldsymbol{s}_i$ 进行编码，而是对 N 个信源符号的序列 $\boldsymbol{s}_i$ 进行编码，然后对 N 求算术平均。

定理 4-5 可视为定理 4-6 在 $N=1$ 时的特殊情况。

【证明】将 S^N 视为一个新的离散无记忆信源，其熵为 $H_r(S^N)$，平均码长为$\bar{L}_N$。

根据定理 4-5 可得

$$H_r(S^N) \leqslant \bar{L}_N < H_r(S^N) + 1$$

由于离散无记忆信源的 N 次扩展信源 S^N 的熵 $H_r(S^N)$ 是信源 S 的熵 $H_r(S)$ 的 N 倍，即 $H_r(S^N) = NH_r(S)$，代入上式得

$$NH_r(S) \leqslant \bar{L}_N < NH_r(S) + 1$$

两边除以 N，得

$$H_r(S) \leqslant \frac{\bar{L}_N}{N} < H_r(S) + \frac{1}{N}$$

当 $N \to \infty$ 时，有 $\lim\limits_{N\to\infty} \frac{\bar{L}_N}{N} = H_r(S)$。

证毕。

定理 4-6 的结论推广到平稳遍历的有记忆信源（如马尔可夫信源），便有

$$\lim_{N\to\infty} \frac{\bar{L}_N}{N} = \frac{H_\infty}{\log r}$$

其中，H_∞ 为有记忆信源的极限熵。

定理 4-6 是香农信息论的主要定理之一。该定理指出，要做到无失真信源编码，每个信源符号平均所需最少的 r 元码符号数就是信源的熵值（以 r 进制单位为信息量单位）。若编码的平均码长小于信源的熵值，则唯一可译码不存在，在译码时必然要带来失真或差错。

同时定理还指出，可以通过对扩展信源进行变长编码，当 $N\to\infty$ 时，平均码长 $\frac{\bar{L}_N}{N}$ 可达到这个极限值。可见，信源的信息熵是无失真信源编码平均码长的极限值，也可以认为信源的信息熵[$H(S)$ 或 H_∞] 是表示每个信源符号平均所需最少的二元码符号数。

我们可以看到，定长码和变长码的平均码长的理论极限是一致的，而且要达到这个极限，也就是平均单个信源符号所需的码符号数最少，所用的方法都是对信源的 N 次扩展信源进行编码，但是变长码与定长码的区别在于，变长码在 N 不需很大时就能达到这个极限，而定长码的 N 值通常会大到设备难以实现，而且定长码在达到这个码长极限时往往还会引入一定的失真，而变长码则不会引入失真。

编码后的信息传输率

$$R=\frac{H(S)}{\bar{L}}\ \frac{\text{比特/信源符号}}{\text{码符号/信源符号}}$$

$$=\frac{H(S)}{\bar{L}}\ \text{比特/码符号}$$

其中，$\bar{L}=\frac{\bar{L}_N}{N}\geqslant\frac{H(S)}{\log r}$。所以，$R\leqslant\log r$。

当平均码长 $\bar{L}$ 达到极限值 $\frac{H(S)}{\log r}$ 时，$R=\log r$，编码后信源的信息传输率等于有 r 个码符号的无噪无损信道的信道容量 C，这时信息传输效率最高。因此，无失真信源编码的实质是对离散信源进行适当的变换，使变换后新的码符号信源（信道的输入）尽可能为等概分布，使新信源的每个码符号平均所含的信息量达到最大，从而使信道的信息传输率 R 达到无噪无损信道的信道容量，实现信源与信道理想的统计匹配。这就是香农第一定理的物理意义。

为了衡量各种编码是否已达到极限情况，我们定义变长码的编码效率的概念。

【定义 4-10】 设对信源 S 进行编码得到的平均码长为 $\bar{L}$，则 $\bar{L}\geqslant H_r(S)$，定义 $\eta=\frac{H_r(S)}{\bar{L}}$ 为编码效率，$\eta\leqslant 1$。

对同一信源来说，码的平均码长 $\bar{L}$ 越短，越接近极限 $H_r(S)$，则信息传输率越高，越接近无噪无损信道的信道容量，这时 η 也越接近于 1，所以用编码效率 η 来衡量各种编码的优劣。

另外，为了衡量各种编码与最佳码的差距，引入码的剩余度的概念。

【定义 4-11】 $\gamma=1-\eta=1-\frac{H_r(S)}{\bar{L}}$ 为码的剩余度。

在二元无噪无损信道中，$r=2$，$\eta=\frac{H(S)}{\bar{L}}$，所以在二元无噪无损信道中信息传输率为

$$R=\frac{H(S)}{\bar{L}}=\eta$$

注意，它们的数值相同，单位不同。η 是个无单位的比值，而 R 的单位是比特/码符号。因此，在二元信道中可直接用码的效率来衡量编码后信道的信息传输率是否提高。当 $\eta=1$

时，即 $R=1$，达到二元无噪无损信道的信道容量，编码效率最高，码剩余度为零。

与定长码一样，通过对扩展信源进行编码，可以提高编码后信道的信息传输率。

【例 4.5】有一离散无记忆信源 $\begin{bmatrix} S \\ P \end{bmatrix}=\begin{bmatrix} s_1 & s_2 \\ \frac{3}{4} & \frac{1}{4} \end{bmatrix}$，求其信息传输率及二次、三次、四次扩展信源的信息传输率。

【解】

信源熵

$$H(S)=\frac{1}{4}\log 4+\frac{3}{4}\log\frac{4}{3}=0.811 \text{ 比特/符号}$$

用二元码符号{0,1}来构造一个即时码 $s_1\to 0$，$s_2\to 1$，这时 $\bar{L}=1$，编码效率 $\eta=\dfrac{H(S)}{\bar{L}}=0.811$。

二元码编码后的信息传输率和编码效率在数值上是相等的，$R=0.811$ 比特/码符号。

如果对信源 S 的二次扩展信源 S^2 进行编码，则 S^2 和它的一种即时码如表 4.8 所示。这时码的平均长度

$$\bar{L}_2=\frac{9}{16}\times 1+\frac{3}{16}\times 2+\frac{3}{16}\times 3+\frac{1}{16}\times 3$$

$$=\frac{27}{16} \text{ 二元码符号/二个信源符号}$$

表 4.8　二次扩展码

s_i	$p(s_i)$	即时码
$s_1\ s_1$	$\frac{9}{16}$	0
$s_1\ s_2$	$\frac{3}{16}$	10
$s_2\ s_1$	$\frac{3}{16}$	110
$s_2 s_2$	$\frac{1}{16}$	111

单个信源符号的平均码长

$$\frac{\bar{L}_2}{2}=\frac{27}{32} \text{ 二元码符号/信源符号}$$

编码效率为

$$\eta_2=\frac{0.811\times 32}{27}=0.961$$

$$R_2=0.961 \text{ 比特/码符号}$$

可见，编码复杂了些，但信息传输效率有了提高。

用同样的方法进一步对信源的三次和四次扩展信源进行编码，并求出其编码效率为

$$\eta_3=0.985$$

$$\eta_4=0.991$$

信道的信息传输率分别为

$$R_3=0.985 \text{ 比特/码符号}$$

$$R_4=0.991 \text{ 比特/码符号}$$

将此例与例 4.3 比较，对于同一信源，要求编码效率达到 96% 时，变长码只需对二次扩展信源（$N=2$）进行编码，而等长码则要求 $N>4.13\times 10^7$。因此用变长码编码时，N 不需很大就可以达到相当高的编码效率，而且可实现无失真编码，随着扩展信源次数 N 的增加，编码效率越来越接近于 1，编码后的信息传输率 R 也越来越接近无噪无损二元信道的信道容量（$C=1$ 比特/码符号），达到信源与信道匹配，使信道得到充分利用。

4.4 变长码的编码方法

本节介绍的编码方法如香农编码、香农－费诺－埃利斯编码、霍夫曼编码、费诺编码均为统计匹配编码，都是对出现概率较高的信源符号编短码，而对出现概率较小的信源符号编长码，从而使平均码长最短，达到最佳编码的目的。

4.4.1 香农编码

香农第一定理指出了平均码长与信源熵之间的关系，也指出了可以通过编码使码长达到极限值。如何构造这种码？香农码的方法是选择每个码字长度 l_i 满足

$$l_i = \left\lceil \log \frac{1}{p(s_i)} \right\rceil \qquad i = 1, 2, \cdots, q$$

式中，$\lceil x \rceil$ 表示不小于 x 的整数，即 x 为整数时等于 x，x 不是整数时，等于 x 取整加 1。

这样选择的码长一定满足 Kraft 不等式，所以一定存在这样的即时码。按照香农编码方法构造的码，其平均码长 $\bar{L}$ 不超过上界，即

$$\bar{L} < H_r(S) + 1$$

只有当信源符号的概率分布满足 $\left(\frac{1}{r}\right)^{l_i}$（$l_i$ 是正整数）形式时，$\bar{L}$ 才能达到极限值 $H_r(S)$。一般情况下，香农编码的 $\bar{L}$ 不是最短的，即编出来的码不是紧致码（最佳码）。

香农编码的具体方法如下：

① 将所有 q 个信源符号按其概率的递减次序排列，即

$$p(s_1) \geqslant p(s_2) \geqslant \cdots \geqslant p(s_q)$$

② 按下式依次计算每个信源符号的二元码码长 l_i：

$$l_i = \left\lceil \log \frac{1}{p(s_i)} \right\rceil \qquad i = 1, 2, \cdots, q$$

③ 计算每个信源符号的累加概率 $F(s_i)$，并变换成二进制小数得到其码字。累加概率

$$F(s_i) = \sum_{k=1}^{i-1} p(s_k) \qquad i = 1, 2, \cdots, q$$

将累加概率 $F(s_i)$ 变换成二进制小数，取小数点后 l_i 位数作为第 i 个信源符号的码字。

【例 4.6】参见表 4.9，按照以上步骤对一有 7 个信源符号的信源编码。例如，当 $i=4$ 时，先求第 4 个信源符号的二元码的码长 $l_4 = \lceil -\log p(s_4) \rceil = 3$，因此码长取为 3。

表 4.9 香农编码

信源符号 s_i	概率 $p(s_i)$	累加概率 $F(s_i)$	$-\log p(s_i)$	码长 l_i	二元码
s_i	0.20	0	2.34	3	000
s_2	0.19	0.2	2.41	3	001
s_3	0.18	0.39	2.48	3	011
s_4	0.17	0.57	2.56	3	100
s_5	0.15	0.74	2.74	3	101
s_6	0.10	0.89	3.34	4	1110
s_7	0.01	0.99	6.66	7	1111110

再计算累加概率：

$$F(s_4)=\sum_{K=1}^{3}p(s_k)=p(s_1)+p(s_2)+p(s_3)=0.57$$

将累加概率 $F(s_4)$ 变成二进制小数：

$$F(s_4)=0.57=0\times2^{0}+1\times2^{-1}+0\times2^{-2}+0\times2^{-3}+1\times2^{-4}+\cdots$$

即

$$F(s_4)=(0.57)_{10}=(0.1001\cdots)_2$$

根据码长 $l_4=3$，取小数点后三位作为第 4 个信源符号的二元码，即“100”。其他信源符号的编码可依次求得，过程不再赘述。

由表 4.9 可以看出，有 5 个三位的二元码，各码字至少有一位码符号不同。这个码是唯一可译码，而且是即时码。

平均码长

$$\bar{L}=\sum_{i=1}^{7}p(s_i)l_i=3.14\ \text{码符号/信源符号}$$

编码后信息传输率

$$R=\frac{H(S)}{\bar{L}}=\frac{2.61}{3.14}=0.831\ \text{比特/码符号}$$

4.4.2 香农－费诺－埃利斯编码

香农－费诺－埃利斯编码与香农编码方法大致相同，但有三点不同，一是它不需要对信源符号概率进行排序，二是它的累加概率是修正的累加概率，三是码长的计算方法不同。

修正的累加概率的计算方法为

$$\bar{F}(s_i)=\sum_{k=1}^{i-1}p(s_k)+\frac{1}{2}p(s_i)\qquad i=1,2,\cdots,q$$

码长的计算方法为

$$l_i=\left\lceil\log\frac{1}{p(s_i)}\right\rceil+1\qquad i=1,2,\cdots,q$$

4.4.3 二元霍夫曼码

二元霍夫曼码是霍夫曼于 1952 年提出的一种构造紧致码的方法。具体如下：

① 将所有 q 个信源符号按其概率的递减次序排列，即

$$p(s_1)\geqslant p(s_2)\geqslant\cdots\geqslant p(s_q)$$

② 合并概率最小的两个信源符号的概率，从而得到只包含 $q-1$ 个符号的新信源 S_1，称为缩减信源。在那两个概率最小的两个信源符号旁边用 0、1 码符号标注。

③ 把缩减信源 S_1 的符号仍按概率大小递减次序排列，再将两个概率最小的信源符号用 0 和 1 码符号标注，并且合并概率，这样又形成了 $q-2$ 个信源符号的缩减信源 S_2。

④ 依次继续，直至信源最后只剩下两个信源符号为止，将这最后两个信源符号分别用二元码符号 0 和 1 标注。

⑤ 从最后一级缩减信源开始，向前回溯，将每次标注的码符号连接起来就得到各信源符号所对应的码符号序列，即相应的码字。

【例 4.7】以例 4.6 信源为例编制二元霍夫曼码（如表 4.10 所示）。

表 4.10　霍夫曼编码

码字	信源符号	编码过程	码长
10	s_1	0.20 — 0.20 — 0.26 — 0.35 — 0.39 — 0.61 (0)	2
11	s_2	0.19 — 0.19 — 0.20 — 0.26 — 0.35 (0) — 0.39 (1)	2
000	s_3	0.18 — 0.18 — 0.19 — 0.20 (0) — 0.26 (1)	3
001	s_4	0.17 — 0.17 — 0.18 (0) — 0.19 (1)	3
010	s_5	0.15 — 0.15 (0) — 0.17 (1)	3
0110	s_6	0.10 (0) — 0.11 (1)	4
0111	s_7	0.01 (1)	4

【解】

该码的平均码长为

$$\bar{L}=\sum_{i=1}^{7}p(s_i)l_i$$

$$=0.2\times2+0.19\times2+0.18\times3+0.17\times3+0.15\times3+0.1\times4+0.01\times4$$

$$=2.72\ \text{比特/码符号}$$

其编码效率为

$$\eta=\frac{H_r(S)}{\bar{L}}=\frac{2.61}{2.72}=0.960$$

从霍夫曼码的编码方法可知，这样得到的码并不是唯一的，原因有两个：

① 在每次对信源缩减时，概率最小的两个信源符号对应的码符号 0 和 1 是可以互换的，所以可得到不同的霍夫曼码。

② 对信源进行缩减时，如果两个概率最小的信源符号合并后的概率与其他信源符号的概率相同，则在缩减信源中进行概率排序的次序可以是任意的，因此会得到不同的霍夫曼码。

表 4.11 给出了同一信源的两种霍夫曼码，它们的平均码长和编码效率都相同，都是紧致码，但是质量不完全相同，因为它们的码长方差不同。

表 4.11　霍夫曼码之间的比较

信源符号 s_i	概率 $p(s_i)$	码 1	码 1 的码长	码 2	码 2 的码长
s_1	0.4	1	1	00	2
s_2	0.2	01	2	10	2
s_3	0.2	000	3	11	2
s_4	0.1	0010	4	010	3
s_5	0.1	0011	4	011	3

平均码长为

$$\bar{L}=\sum_{i=1}^{5}p(s_i)l_i=2.2\ \text{码符号/信源符号}$$

编码效率为

$$\eta=\frac{H_r(S)}{\bar{L}}=0.965$$

码 1 的码长方差为

$$\begin{aligned}\sigma_1^2&=\sum_{i=1}^{5}p(s_i)(l_i-\bar{L})^2\\&=0.4\times(1-2.2)^2+0.2\times(2-2.2)^2+0.2\times(3-2.2)^2+\\&\quad 0.1\times(4-2.2)^20.4\times(0.4-2.2)^2\\&=1.36\end{aligned}$$

码 2 的码长方差为

$$\begin{aligned}\sigma_2^2&=\sum_{i=1}^{5}p(s_i)(l_i-\bar{L})^2\\&=0.4\times(2-2.2)^2+0.2\times(2-2.2)^2+0.2\times(2-2.2)^2+\\&\quad 0.1\times(3-2.2)^2+0.1\times(3-2.2)^2\\&=0.16\end{aligned}$$

由此可见，码 2 的码长方差要比码 1 的码长方差小很多，因此码 2 的码长更均匀，质量更好。

从此例可以看出，霍夫曼编码在信源缩减排列时，应使合并的信源符号位于缩减信源中尽可能高的位置上，这样可以使合并的信源符号码长较短，充分利用短码，而非合并的信源符号码长较长，所以得到方差最小的码。

霍夫曼码的编码过程也可以用树图来表示。上例中，信源的霍夫曼码的编制过程可以用图 4.4 表示。可以看出霍夫曼码是即时码。

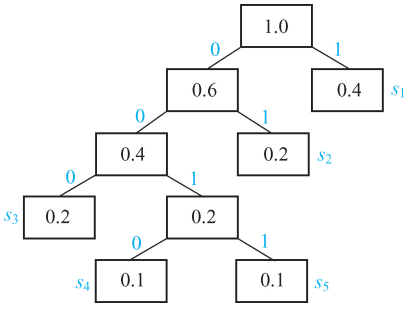


图 4.4　霍夫曼码的树图

霍夫曼码用概率匹配的方法进行信源编码，它有两个明显特点：① 霍夫曼码的编码方法保证了概率大的信源符号对应于短码，概率小的信源符号对应于长码，可充分利用短码；② 每次缩减信源的最后两个码字有相同的码长，并且总是最后一位码元不同，前面各位码元相同。这两个特点保证了所得的霍夫曼码一定是最佳码。

【定理 4-7】霍夫曼码是紧致码。

【证明】在这里证明二元霍夫曼码是紧致码，其结论可以推广到 r 元霍夫曼码。

由于霍夫曼码最后一步得到的缩减信源只有两个信源符号，编码为 0 和 1，它们是紧致码，所以我们可以假设缩减后的编码是紧致码，然后证明缩减前的编码也是紧致码，这样就可以证明最后得到的霍夫曼码是紧致码。

设霍夫曼码中第 j 步缩减信源为 S_j。S_j 有 m 个信源符号，被编码为 C_j，其平均码长为 $\bar{L}_j$，则

$$\bar{L}_j = \sum_{i=1}^{m} p(s_i) l_i$$

假设 S_j 中的某一元素 s_m 由前一次缩减信源 S_{j-1} 中的两个概率最小的信源符号 s_{m_0} 和 s_{m_1} 合并而来，即

$$p(s_m) = p(s_{m_0}) + p(s_{m_1})$$

设 C_{j-1} 为第 $j-1$ 步缩减信源 S_{j-1}（S_{j-1} 有 $m+1$ 个元素）的编码，其平均码长为 $\bar{L}_{j-1}$：

$$\begin{aligned}\bar{L}_{j-1} &= \sum_{i=1}^{m-1} p(s_i) l_i + p(s_{m_0})(l_m + 1) + p(s_{m_1})(l_m + 1) \\ &= \sum_{i=1}^{m} p(s_i) l_i + p(s_{m_0}) + p(s_{m_1}) \\ &= \bar{L}_j + p(s_{m_0}) + p(s_{m_1})\end{aligned}$$

缩减信源 S_j 和 S_{j-1} 的平均码长之差是一个与码长 l_i 无关的固定常数 $p(s_{m_0}) + p(s_{m_1})$，所以如果平均码长 $\bar{L}_j$ 最小，则 $\bar{L}_{j-1}$ 也最小。也就是说，如果 C_j 是缩减后信源 S_j 的紧致码，则 C_{j-1} 是缩减前信源 S_{j-1} 的紧致码。

由于最后一级缩减信源可以肯定是紧致码，则由霍夫曼编码方法使得它前面一级缩减信源的编码也一定是紧致码，再前面一级缩减信源的编码也是紧致码。由递推关系可知，信源 S 所得的编码是紧致码。

4.4.4 r 元霍夫曼码

二进制霍夫曼码的编码方法很容易推广到 r 进制的情形，只是编码过程中构成缩减信源时，每次都是将 r 个概率最小的信源符号合并，并分别用 $0,1,\cdots,r-1$ 码符号表示。

为了充分利用短码，使霍夫曼码的平均码长最短，必须使最后一个缩减信源有 r 个信源符号。因此，对于 r 元霍夫曼编码，信源 S 的符号个数 q 必须满足 $q=(r-1)\theta+r$。θ 表示信源缩减次数，如果不满足上式，则可以在最后增补一些概率为 0 的信源符号，因此上式又可以写成 $q+i=(r-1)\theta+r$。i 为增加的信源符号个数，是满足上式的最小正整数或 0。对于 $r=2$ 时的二元码，信源 S 的符号个数 q 必定满足 $q=\theta+2$。

【例 4.8】构造一个三元霍夫曼码。

【解】

编码结果如表 4.12 所示，满足 $q+i=(r-1)\theta+r$ 的 i 的最小值为 $i=0$，所以不需增加概率为 0 的信源符号。

表 4.12　三元霍夫曼编码

信源符号	概率 $p(s_i)$	码字	码长
s_1	0.4	1	1
s_2	0.2	2	1
s_3	0.2	00	2
s_4	0.1	01	2
s_5	0.1	02	2

该码的平均码长为

$$\begin{aligned}\bar{L} &= \sum_{i=1}^{5} p(s_i) l_i \\ &= 0.4 \times 1 + 0.2 \times 1 + 0.2 \times 2 + 0.1 \times 2 + 0.1 \times 2 \\ &= 1.4 \text{ 三进制码符号/信源符号}\end{aligned}$$

信息传输率为

$$R = \frac{H(S)}{\bar{L}} = \frac{2.122}{1.4} = 1.515 \text{ 比特/三进制码符号}$$

编码效率为

$$\eta = \frac{H_3(S)}{\bar{L}} = \frac{1.339}{1.4} = 0.956$$

信源的 N 次扩展信源同样可以使用霍夫曼编码方法。由于霍夫曼码是紧致码，所以编码后单个信源符号平均码长随 N 的增加很快接近于极限值——信源熵。

4.4.5 费诺码

费诺码与香农码一样，也不是最佳的编码方法，但是某些情况下也能得到紧致码。

费诺码编码过程如下：

① 将信源符号 $s_i(i=1,2,\cdots,q)$ 以概率递减次序排列，即

$$p(s_1) \geqslant p(s_2) \geqslant \cdots \geqslant p(s_q)$$

② 将依次排列的信源符号分为两组，使两组的概率和基本相等，并对各组赋予二元码符号 0 和 1。

③ 将每大组的信源符号进一步再分成两组，使划分后的两组的概率和近似相等，又分别赋予各组二元码符号“0”和“1”。

④ 如此重复，直至每组只剩下一个信源符号为止。

⑤ 信源符号所对应的码符号序列即为费诺码。

【例 4.9】信源与例 4.6 和例 4.7 的相同，请编制费诺码。

【解】

编码过程如表 4.13 所示。

表 4.13 费诺码

信源符号	概率	第 1 次分组	第 2 次分组	第 3 次分组	第 4 次分组	二元码	码长
s_1	0.20	0	0			00	2
s_2	0.19		1	0		010	3
s_3	0.18			1		011	3
s_4	0.17	1	0			10	2
s_5	0.15		1	0		110	3
s_6	0.10			1	0	1110	4
s_7	0.01				1	1111	4

该码的平均码长为

$$\bar{L}=\sum_{i=1}^{7}p(s_i)l_i$$
$$=0.20\times2+0.19\times3+0.18\times3+0.17\times2+$$
$$0.15\times3+0.10\times4+0.01\times4$$
$$=2.74\ \text{码符号/信源符号}$$

信息传输率为

$$R=\frac{H(s)}{\bar{L}}=\frac{2.61}{2.74}=0.953\ \text{比特/码符号}$$

对同一信源三种编码方法的性能比较如表 4.14 所示。

表 4.14　三种编码方法的比较

编码	平均码长 $\bar{L}$	信息传输率 R
香农码	3.14	0.831
霍夫曼	2.72	0.959
费诺码	2.74	0.953

如果信源概率满足

$$p(s_i)=\left(\frac{1}{r}\right)^{l_i}\qquad i=1,2,\cdots,q$$

l_i 为正整数，则三种码都能得到紧致码。

上面讲述的几种编码都是针对离散无记忆信源的单个信源符号的编码，因此在编码时未考虑其信源符号之间的相关性，仅考虑了信源符号概率分布的不均匀性。对于有记忆信源，需要考虑信源符号间的相关性，对符号序列进行编码，才能进一步提高编码效率。

以上讨论了几种常用的编码方法，并且证明了霍夫曼码是最佳码，当 N 不是很大时，它能使无失真编码的效率接近于 1，但在实际使用时霍夫曼编码的设备还是比较复杂的。

首先，由于每个信源符号所对应的码长不同，一般情况下，信源符号以匀速输出，信道也是匀速传输的。通过霍夫曼编码后，有长码也有短码，会造成编码输出的信息速率不是常量，因此不能直接由信道来传输。为了适应信道，必须增加缓冲寄存器，将编码输出暂存在缓冲器中，再匀速向信道输出。但当缓冲器容量有限时，会出现缓冲器溢出或取空的现象。比如，当信源连续输出低概率的信源符号时，因为码长较长，有可能使缓冲器存不下而溢出；反之，当信源连续输出高概率符号时，则有可能使缓冲区输入比特数小于向信道中输出的比特数，以致缓冲器取空。所以，一般变长码只适用于有限长的信息传输，或者在输出一批消息后能暂停一下。

其次，存在差错扩散的问题。变长码的一个码元的差错可能被误认为是另一个码字而点断，结果后面一系列码字也可能会译错。为了避免出现这种情况，需要进行纠错编码或出错重发等，这样会降低信息传输的有效性。

最后，霍夫曼码的编译码需要用查表的方法来进行。在信息传输过程中必须先存储和传输该霍夫曼编码表，这会影响信息的传输效率。特别是当 N 增大时，所需存储的码表也增大，所以使得设备复杂化，且查表搜索时间增大。根据信源的统计特性，还需要事先建立霍夫曼编码表。因此，这种方法只适用于已知其统计特性的信源。如果信源的统计特性不知道或经常发生变化，就要采用所谓的通用编码方法来解决，如后面要介绍的 LZW 编码。

尽管如此，霍夫曼编码方法还是一种有效的无失真信源编码方法，便于硬件实现和计算机软件实现，适用于文件传真、语声处理和图像处理的数据压缩。

4.5 实用的无失真信源编码方法

4.5.1 游程编码

对于某些相同符号会连续出现的信源，特别是对于二值图像文件（如图 4.5 所示），采用游程编码会得到较高的编码效率。游程编码又称为“运行长度编码”或“行程编码”，仍然是一种统计编码，属于无失真信源编码。游程编码算法简单，速度快，编码效率与信源符号序列中字符重复出现的概率及长度有关，字符重复出现的次数越多，重复字符串的平均长度越长，编码效率越高。

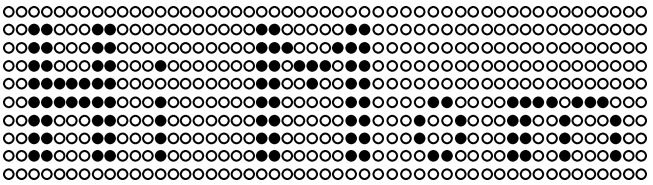


图 4.5 二值图像

游程编码的基本原理是：用一个符号值或串长表示连续出现的信源符号（这些连续出现的信源符号构成了一段连续的“游程”，游程编码因此而得名），使码符号序列长度小于原始信源符号序列的长度。基本的游程编码的数据结构如图 4.6 所示。

符号码	标识码	游程长度

图 4.6 游程编码的数据结构

例如，要表示 BBBBBBBBBBXXXXXXXXXAAAAAAUUUUUUUUUUUU，游程编码为“B#10X#9A#6U#13”，#是游程编码的标识符，表示有一个字符串在此位置，它前面是构成串的符号，后面是游程长度。可见，游程编码的码符号序列长度远远小于原始信源符号序列的长度。

对于黑、白二值文件来说，以上方法还有可以改进的地方。

① 黑白游程总是交替出现，可以规定第一游程为白游程，否则可以将该白游程的长度定为 0，这样可以省略符号码和标识符，只需对游程长度进行编码。

② 不同游程长度出现的概率不同，对游程长度进行编码采用霍夫曼编码，概率大的编长码，概率小的编短码。

修正的霍夫曼编码（Modified Huffman，MH）是一种实用的游程编码方案，是 CCITT 为三类传真机制定的图像编码方式。霍夫曼编码的码表是由各类传真文件（打字文件、电路图、手写文稿、气象图等）的统计特性得到的，并且固定不变，因而在多数情况下，并非紧致码。

为了保证收发图文颜色同步，每行总是从白色游程开始（如第一像素为黑色，则此长度可设为零）。对概率不同的游程长度编码码字不同，如长度为 2 和 3 的黑游程的码字分别为 11 和 10，而长度为 63 的黑游程的码字为 000001100111。

如果游程长度超过 63 个像素，则码字分成两部分，前为组合基干码，后为结尾码，这样可以提高编码效率。比如，长度为 73 的白游程的码字为 1101110100，其中前面的 11011 为组合基干码，后面的 10100 为结尾码。

其基本的编码规范如下：

- 游程长度在 0～63 时，直接查表用相应的结尾码作为码字。

- 游程长度在 64 ~ 1728（标准 A4 大小的黑白传真文件每行 1728 像素）范围内时，用组合码加上结尾码作为相应的码字。
- 每行的第一个游程规定为白游程（长度可为零），每行用一个结束码（EOL）终止。
- 在传输时，每页数据之前加一个结束码，每页尾部连续使用 6 个结束码。

【例 4. 10】某行经过霍夫曼编码压缩的传真信号为

001101010101101010111101100001100110000000000001

请恢复黑白像素序列，并计算压缩比。

【解】码字为

00110101　010　110101　011　11011 + 000011　0011　000000000001（结束码）

查表 4. 15 和表 4. 16 可知，原来的信号是：0 白 1 黑 15 白 4 黑 77 白 5 黑，压缩比为$\frac{1728}{47}$ = 36. 7∶1。

表 4. 15　结尾码码表

游 程 长 度	白游程编码	黑游程编码	游 程 长 度	白游程编码	黑游程编码
0	00110101	0000110111	32	00011011	000001101010
1	000111	010	33	00010010	000001101011
2	0111	11	34	00010011	000011010010
3	1000	10	35	00010100	000011010011
4	1011	011	36	00010101	000011010100
5	1100	0011	37	00010110	000011010101
6	1110	0010	38	00010111	000011010110
7	1111	00011	39	00101000	000011010111
8	10011	000101	40	00101001	000001101100
9	10100	000100	41	00101010	000001101101
10	00111	0000100	42	00101011	000011011010
11	01000	0000101	43	00101100	000011011011
12	001000	0000111	44	00101101	000001010100
13	000011	00000100	45	00000100	000001010101
14	110100	00000111	46	00000101	000001010110
15	110101	000011000	47	00001010	000001010111
16	101010	0000010111	48	00001011	000001100100
17	101011	0000011000	49	01010010	000001100101
18	0100111	0000001000	50	01010011	000001010010
19	0001100	00001100111	51	01010100	000001010011
20	0001000	00001101000	52	01010101	000000100100
21	0010111	00001101100	53	00100100	000000110111
22	0000011	00000110111	54	00100101	000000111000
23	0000100	00000101000	55	01011000	000000100111
24	0101000	00000010111	56	01011001	000000101000
25	0101011	00000011000	57	01011010	000001011000
26	0010011	000011001010	58	01011011	000001011001
27	0100100	000011001011	59	01001010	000000101011
28	0011000	000011001100	60	01001011	000000101100
29	00000010	000011001101	61	00110010	000001011010
30	00000011	000001101000	62	00110011	000001100110
31	00011010	000001101001	63	00110100	000001100111

表 4.16　组合基干码码表

游程长度	白游程编码	黑游程编码	游程长度	白游程编码	黑游程编码
64	11011	0000001111	960	011010100	0000001110011
128	10010	000011001000	1024	011010101	0000001110100
192	010111	000011001001	1088	011010110	0000001110101
256	0110111	000001011011	1152	011010111	0000001110110
320	00110110	000000110011	1216	011011000	0000001110111
384	00110111	000000110100	1280	011011001	0000001010010
448	01100100	000000110101	1344	011011010	0000001010011
512	01100101	0000001101100	1408	011011011	0000001010100
576	01101000	0000001101101	1472	010011000	0000001010101
640	01100111	0000001001010	1536	010011001	0000001011010
704	011001100	0000001001011	1600	010011010	0000001011011
768	011001101	0000001001100	1664	011000	0000001100100
832	011010010	0000001001101	1728	010011011	0000001100101
896	011010011	0000001110010	—	—	—

4.5.2　算术编码

霍夫曼码虽然是个实用、高效的编码方法，但对于信源符号个数不多且概率分布比较均匀的信源，需要对较长的信源符号序列进行编码才能得到较高的编码效率（见例 4.5），所以需要预先计算信源符号序列的概率分布。

与霍夫曼码不同，算术编码是一种非分组码，不需要计算出所有 N 长信源序列的概率分布及码表，可以直接对输入的信源符号序列编码输出。这种方法是由香农 – 费诺 – 埃利斯编码直接推广得到的。

香农 – 费诺 – 埃利斯编码将累加概率分布函数的[0,1]区间分成许多互不重叠的小区间，每个信源符号对应于每个小区间，每个区间的长度等于这个信源符号的概率，在此区间内取一点，取该点二进制小数点后若干位作为这个信源符号的码字，码字的长度由这个信源符号的概率决定。

算术编码的思想与之相同，不过算术编码不是针对单个的信源符号，而是针对整个信源符号序列。或者说，其他编码方法是把输入的消息分割为符号，然后对每个符号进行编码，而算术编码是直接把整个输入的消息编码为一个数，一个在 0 到 1 之间的小数。因此，算术编码算法的关键仍然是计算 N 长信源符号序列 $\boldsymbol{s}$ 的概率 $p(\boldsymbol{s})$ 和累加概率 $F(\boldsymbol{s})$，然后用区间 $[F(\boldsymbol{s}),F(\boldsymbol{s})+p(\boldsymbol{s}))$ 中的一个值来作为 $\boldsymbol{s}$ 的码字。计算的过程是随着信源符号序列变长，采用递推的方法来完成的。信源符号序列的概率决定编码的效率，概率越小，就需要更多的码符号来表示这个信源符号序列。

例如，信源符号集 $S=\{a_1,a_2,a_3,a_4\}$，$p(a_1)=0.6$，$p(a_2)=0.2$，$p(a_3)=0.1$，$p(a_4)=0.1$，对符号序列 $a_1a_3a_4$ 编码。先确定第一个符号 a_1 的概率区间，如图 4.7(a)中的粗线所示，再确定前两个符号 a_1a_3 的概率区间，如图 4.7(b)中的粗线所示，最后确定整个符号序列 $a_1a_3a_4$ 的概率区间，如图 4.7(c)中的粗线所示。

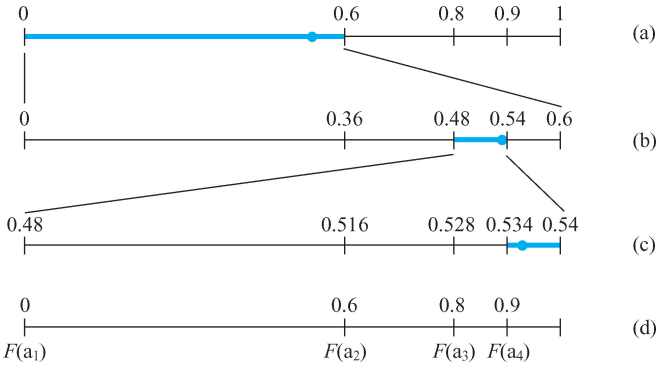


图 4.7　算术编码

假设已知 N 长信源符号序列 $\boldsymbol{s}$ 的概率 $p(\boldsymbol{s})$ 和累加概率 $F(\boldsymbol{s})$，由递推关系很容易得到 $N+1$ 长信源符号序列 $\boldsymbol{sr}$ 的概率 $p(\boldsymbol{sr})$ 和累加概率 $F(\boldsymbol{sr})$，r 为新添加的符号：

$$\begin{cases} F(\boldsymbol{sr}) = F(\boldsymbol{s}) + p(\boldsymbol{s})F(r) \\ p(\boldsymbol{sr}) = p(\boldsymbol{s})p(r) \end{cases}$$

根据下式计算信源符号序列 $\boldsymbol{s}$ 的码长：

$$l = \left\lceil \log \frac{1}{p(\boldsymbol{s})} \right\rceil$$

将信源符号序列 $\boldsymbol{s}$ 的累加概率 $F(\boldsymbol{s})$ 变换成二进制小数，取 l 位为码字。注意，如果 l 位后还有尾数，则需要进位。

符号序列 $a_1a_3a_4$ 的概率为

$$\begin{aligned} p(a_1a_3a_4) &= p(a_1) \times p(a_3) \times p(a_4) \\ &= 0.6 \times 0.1 \times 0.1 = 0.006 \end{aligned}$$

累加概率为

$$\begin{aligned} F(a_1a_3a_4) &= F(a_1) + p(a_1) \times F(a_3) + p(a_1a_3) \times F(a_4) \\ &= 0 + 0.6 \times 0.8 + 0.6 \times 0.1 \times 0.9 = 0.534 \end{aligned}$$

码长为

$$l = \left\lceil \log \frac{1}{0.006} \right\rceil = 8$$

将累加概率变换成二进制小数：

$$F(a_1a_3a_4) = (0.534)_{10} = (0.100010001\cdots)_2$$

所以编码码字为 10001001。注意最后一位有进位。这样得到的码字代表的数值一定在区间 $[F(\boldsymbol{s}), F(\boldsymbol{s}) + p(\boldsymbol{s}))$ 内，即图 4.7(a) ~ (c) 中数轴上的黑点所表示的值。对方收到 10001001 后，根据该码字表示的区间位置，可以判断出第一个符号为 a_1，因为它在区间 $[0, 0.6)$ 内，第二个符号为 a_3，因为它在区间 $[0.48, 0.54)$ 内。以此类推，对方很容易解出原符号序列为 $a_1a_3a_4$。

因此，算术编码是从全序列出发，采用递推形式的连续编码。它不是将单个信源符号映射成一个码字，而是将整个输入符号序列映射为实数轴上 $[0,1)$ 区间内的一个小区间，区间长度等于该序列的概率；再在该小区间内选择一个代表性的二进制小数，作为实际的编码输

出，从而达到高效编码的目的。不论是否为二元信源，也不论数据的概率分布如何，其平均码长均能逼近信源的熵。信源符号概率接近时，建议使用算术编码，这种情况下的效率高于霍夫曼编码。

在算术编码中需要注意几个问题：

① 由于实际计算机的精度不可能无限长，运算中很容易出现溢出的问题。

② 算术编码器对整个消息序列只产生一个码字，这个码字是在间隔[0,1)上的一个实数，因此，译码器在接收到表示这个实数的所有消息序列之前不能进行译码。

③ 算术编码是一种对错误很敏感的编码方法，如果有一位发生错误，就会导致整个消息译错。

算术编码可以是静态的或自适应的。在静态算术编码中，信源符号的概率是固定的。在自适应算术编码中，信源符号的概率根据编码时符号出现的频率动态地进行修改，在编码期间估算信源符号的概率。需要开发动态算术编码的原因是，事前知道精确的信源概率很难，且不切实际。

算术编码方法比霍夫曼编码等熵编码方法要复杂，但它不需要传输像霍夫曼编码的码表，同时算术编码还有自适应能力，所以算术编码是一种实现高效数据压缩的编码方法。这也是我们学习算术编码的意义。

4.5.3 LZW 编码

对于统计特性已知的平稳信源，霍夫曼码和算术码的编码效率已非常高，而且实现也不算太困难，但在信源统计特性不可知时，就需要用到具有自适应性能的通用编码方法。LZW 编码是一种高效的通用编码。LZW 编码是一种基于字典的编码方法，继承了 LZ77 和 LZ78 压缩效果好、速度快的优点，而且在算法描述上更容易被人们接受，实现也相对简单。其后发展出来的各式各样的字典编码算法，基本上是这三种编码算法的分支或变体。这三种编码算法现在已成为计算机文件压缩的标准算法。

基于字典的编码方法既没有高深的理论背景，也没有复杂的数学公式，它们只是巧妙地将字典方法运用于编码技术。字典方法就是用字典中的页码和行号代替文章中的每个单词。而且 LZW 编码是一种自适应编码，它的字典是直接由被压缩文件在编码过程中生成的。

1. 基于字典的编码方法的基本原理

计算机文件是以字节为单位组成的。每字节的取值范围为 0 到 255。每字节都视为一个字符，共 256 种字符。再把连续的若干“字符”视为一个“单词”，这样全部字符（可视为单字节单词）、单词及它们对应的序号（码字）组成字典。编码时，把字符、单词用对应的码字来代替。通常字典的容量为 4096，所以序号（码字）的长度为 12 比特，一般“单词”的平均长度远大于 12 比特，因而可以达到压缩的目的。每个单词的序号都相同，因此它是定长码。

例如，在一个文件中，有一个片段的内容为“ABCD 空空空空空 20000”共 14 字节，长度为 112 比特，编码时被分割成 4 个单词，即“ABCD”、“空空空空空”、“2”和“0000”，编码长度为 $4 \times 12 = 48$ 比特，则压缩比为 2.33。

2. 字典的构成

假定字典容量为 4096（0 ~ 4095），序号用 12 比特表示。最后一个单词（第 4095 个单词）为空。单词由前缀字符串和尾字符串两部分组成。其中前缀字符串是字典中已经存在的某个“单词”，用序号表示，尾字符是本单词的最后一个字符。比如，假定 AB 这个单词在字典中已经存在，且序号为 100，那么单词 ABC 的内容可以用 3 字节表示，其中前 2 字节表示前缀单词 AB 的序号 100，最后字节为尾字符 C。将压缩文件恢复为原始文件（解压缩）时，根据单词的这种格式，也是使用递归算法来恢复单词的内容。

3. 编码算法

① 字典初始化：把所有使用到的单字节单词放入字典。为了能压缩任何类型的文件，可以将字典的前 256 个位置（0x000 ~ 0x0FF）依次分配给 0x00 ~ 0xFF 的 256 个单字节单词。这 256 个单字节的单词前缀均为 4095（0xFFF），表示前缀为空。

② 动态数据初始化：初始化新单词存放位置指针 P，将它指向字典的第一个空位置。例如 $P=256$（即 0x100）。读入被压缩文件的第一个字符，作为待处理的当前单词 W 的尾字符。当前单词的前缀为空，即为 0xFFF。

③ 如果文件再没有字符了，输出当前单词 W 的序号，编码结束。如果文件中还有字符，把当前单词 W（的序号）作为前缀，再从被压缩文件中读入一个字符，作为尾字符，得到一个新单词 W_1。

④ 如果字典中已有 W_1，则将 W_1 视为当前单词 W，返回步骤③。

如果字典中没有 W_1（发现一个新单词），先将原单词 W 的序号输出，再将新单词 W_1 增加到字典中，然后把刚刚读入的字符作为当前单词 W，返回步骤③。

4. 解码算法

① 字典初始化：将字典的前 256 个位置（0x000 ~ 0x0FF）依次分配给 0x00 ~ 0xFF 这 256 个单字节单词。

② 动态数据初始化：初始化新单词存放位置指针 P，将它指向字典的第一个空位置，如 $P=256$（即 0x100），读入压缩文件的第一个码字。由于第一个码字必定是一个单字节单词，可以从初始字典中查表得到，译码输出，并记忆它的码字。

③ 如果压缩中已经没有码字，解码结束，否则继续读入一个码字。

④ 如果读入的码字是无效码字（即 0xFFF），则解码结束，否则进入下一步。

⑤ 如果在字典中已经有该码字对应的单词，则采用递归算法，输出该单词的内容。将单词的第一个有效字符作为尾字符，将已经记忆的前一个码字作为前缀，组成一个新单词，写入字典中，然后将当前码字记忆下来，返回步骤③。否则，首先在字典中生成新的单词，再输出这个单词。将新单词的码字记忆下来。返回步骤③，这时的新单词一定是首尾相同的单词。

LZW 编码对于非平稳信源具有较好的处理效果；对于平滑且噪声小的信源具有较高的压缩比，且压缩解压缩速度快；对于数据流中连续重复出现的字节和字串，具有很高的压缩比。LZW 编码常用于图像数据的压缩处理和文本程序等领域的数据压缩。

LZW 编码不适合小文件的压缩（因为压缩编码初期，由于字典中的单词很少，字典对压缩效果的贡献也很少，主要是进行字典的扩充），也不适合太大文件的压缩（因字典容量有限，文件太大时字典满了，效率将受到制约），适合内容有明显单词结构的文件（如文本文件、程序文件）。

为了进一步提高压缩效果并适应超大型文件的压缩需要，LZW 压缩算法不断被改进，如字典的大小根据需要可以扩充，码字的长度也可以不断调整。PKZIP、ART、ARC、LHA、WINZIP 等压缩软件都是各自采用了不同的技术改进而成的。

【例 4.11】信源符号序列为 ABCABDABCAAAABBBABCABCA，其编解码过程如表 4.17和表 4.18 所示。

表 4.17　LZW 码的编码过程

读入字符	要查找的新单词 W_1	当前单词 W	输出码字	字典扩充内容及序号
A				
B	AB	A	041	AB(0x100)
C	BC	B	042	BC(0x101)
A	CA	C	043	CA(0x102)
B	AB			
D	ABD	AB	100	ABD(0x103)
A	DA	D	044	DA(0x104)
B	AB			
C	ABC	AB	100	ABC(0x105)
A	CA			
A	CAA	CA	102	CAA(0x106)
A	AA	A	041	AA(0x107)
A	AA			
B	AAB	AA	107	AAB(0x108)
B	BB	B	042	BB(0x109)
B	BB			
A	BBA	BB	109	BBA(0x10A)
B	AB			
C	ABC			
A	ABCA	ABC	105	ABCA(0x10B)
B	AB			
C	ABC			
A	ABCA			
		ABCA	10B	

表 4.18　LZW 码的译码过程

读入码字	解码输出单词	记忆码字	字典扩充内容及序号
041	A	—	—
042	B	041	AB(0x100)
043	C	042	BC(0x101)
100	AB	043	CA(0x102)
044	D	100	ABD(0x103)
100	AB	044	DA(0x104)
102	CA	100	ABC(0x105)
041	A	102	CAA(0x106)
107	AA	041	AA(0x107)
042	B	107	AAB(0x108)
109	BB	042	BB(0x109)
105	ABC	109	BBA(0x10A)
10B	ABCA	105	ABCA(0x10B)
FFF	—	—	—

扩展阅读：渐进等分割性和典型序列

严格论证定长信源编码定理需要介绍渐进等分割性和 ε 典型序列。在信息论的定理证明中，它是一种重要的数学工具。

当随机试验次数很大时，事件发生的频率具有稳定性。如多次抛掷硬币，出现正面或反面的次数是不定的，但是随着试验次数的增加，出现正面或反面的频率逐渐将稳定于 0.5，这就是随机事件的统计规律性。

对于独立、等同分布的随机变量 $X_1, X_2, X_3, \cdots, X_N$，只要 N 足够大，其算术平均值 $\frac{1}{N}\sum_{i=1}^{N} X_i$ 接近其数学期望值 $E(X)$，即

$$\lim_{N\to\infty} P\left\{ \left| \frac{1}{N}\sum_{i=1}^{N} X_i - E(X) \right| < \varepsilon \right\} = 1$$

也就是说，其算术平均值依概率收敛于数学期望。当 N 很大时，其算术平均值将几乎变成一个常数 $E(X)$，这就是大数定理。

把 $\frac{1}{N}\sum_{i=1}^{N} X_i$ 视为一个随机变量，则

$$E\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = E(X)$$

$$D\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = \frac{\sigma^2}{N\varepsilon^2}$$

根据契比雪夫不等式可以推出，对于独立同分布的随机变量 $X_1, X_2, X_3, \cdots, X_N$，它们具

有相同的数学期望和方差，对于任意正数 ε，有不等式

$$P\left\{\left|\frac{1}{N}\sum_{i=1}^{N}X_i - E(X)\right| \geqslant \varepsilon\right\} \leqslant \frac{\sigma^2}{N\varepsilon^2}$$

和

$$P\left\{\left|\frac{1}{N}\sum_{i=1}^{N}X_i - E(X)\right| \leqslant \varepsilon\right\} \geqslant 1 - \frac{\sigma^2}{N\varepsilon^2}$$

成立，其中 σ^2 为随机变量 $X_1, X_2, X_3, \cdots, X_N$ 的方差。

考虑一个离散无记忆信源 $\begin{bmatrix} S \\ P(S) \end{bmatrix} = \begin{bmatrix} s_1 & \cdots & s_i & \cdots & s_q \\ p(s_1) & \cdots & p(s_i) & \cdots & p(s_q) \end{bmatrix}$ 的 N 次扩展信源

$$\begin{bmatrix} \boldsymbol{S} \\ P(\boldsymbol{S}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{s}_1 & \cdots & \boldsymbol{s}_j & \cdots & \boldsymbol{s}_{q^N} \\ p(\boldsymbol{s}_1) & \cdots & p(\boldsymbol{s}_j) & \cdots & p(\boldsymbol{s}_{q^N}) \end{bmatrix}$$

其中 $S = S_1S_2S_3\cdots S_N$ 是 N 维随机矢量，而 $\boldsymbol{s}_j = s_{j_1}s_{j_2}\cdots s_{j_N}$，$s_{j_1}, s_{j_2}, \cdots, s_{j_N} \in \{s_1, \cdots, s_i, \cdots, s_q\}$。因为是离散无记忆信源的扩展信源，所以

$$\begin{aligned} p(\boldsymbol{s}_j) &= p(s_{j_1})p(s_{j_2})\cdots p(s_{j_N}) \\ &= \prod_{k=1}^{N} p(s_{j_k}) \end{aligned}$$

$$\begin{aligned} I(\boldsymbol{s}_j) &= -\log p(\boldsymbol{s}_j) = -\log\left[\prod_{k=1}^{N} p(s_{j_k})\right] \\ &= -\sum_{k=1}^{N}\log p(s_{j_k}) = \sum_{k=1}^{N} I(s_{j_k}) \end{aligned}$$

式中，$I(\boldsymbol{s}_j)$ 是一个随机变量，其数学期望就是 S 的熵：

$$E[I(s_j)] = H(S) = \sum_{k=1}^{N} E[I(s_{j_k})] = NH(S)$$

$$D[I(\boldsymbol{s}_j)] = N \cdot D[I(s_i)]$$

因为 $D[I(s_i)] < \infty$，所以当 q 为有限值时，$D[I(s_j)] < \infty$。

相互统计独立的随机变量的函数也是相互统计独立的随机变量，所以 $S_1, S_2, S_3, \cdots, S_N$ 是相互统计独立且服从同一概率分布的随机变量，可以推出其自信息 $I(s_{jk})(k = 1, 2, \cdots, N)$ 也是相互统计独立且服从同一概率分布的随机变量：

$$\frac{I(\boldsymbol{s}_j)}{N} = \frac{1}{N}\sum_{k=1}^{N} I(s_{j_k})$$

$$E\left[\frac{I(\boldsymbol{s}_j)}{N}\right] = \frac{1}{N}H(\boldsymbol{S}) = \frac{1}{N}\sum_{k=1}^{N} E[I(\boldsymbol{s}_{j_k})] = H(\boldsymbol{S})$$

所以，$\frac{I(\boldsymbol{s}_j)}{N}$ 依概率收敛于 $H(\boldsymbol{S})$（大数定理），这称为渐进等分割性。

离散无记忆信源的 N 次扩展信源，N 维随机矢量中每一维随机变量相互独立，当序列长度 N 变得很大时，由于统计规律性，N 个随机变量的算术平均将变成一个常数（随机变量的数学期望），也就是 N 维随机矢量中平均每一维随机变量的自信息非常接近于单符号信源熵。因为 $D\left[\frac{I(\boldsymbol{s}_j)}{N}\right] = \frac{D[I(s_i)]}{N}$，根据契比雪夫定理，有以下不等式成立：

$$P\left\{\left|\frac{I(\boldsymbol{s}_j)}{N}-H(S)\right|\geqslant\varepsilon\right\}\leqslant\frac{D[I(s_i)]}{N\varepsilon^2}$$

$$P\left\{\left|\frac{I(\boldsymbol{s}_j)}{N}-H(S)\right|\leqslant\varepsilon\right\}\geqslant1-\frac{D[I(s_i)]}{N\varepsilon^2}$$

令$\frac{D[I(s_i)]}{N\varepsilon^2}=\delta(N,\varepsilon)$，可知$\lim\limits_{N\to\infty}\delta(N,\varepsilon)=0$。这样，可把扩展信源输出的 N 长序列，分成两个集合 G_ε 和$\overline{G}_\varepsilon$：

$$G_\varepsilon=\left\{\boldsymbol{s}_j:\left|\frac{I(\boldsymbol{s}_j)}{N}-H(S)\right|\leqslant\varepsilon\right\}$$

$$\overline{G}_\varepsilon=\left\{\boldsymbol{s}_j:\left|\frac{I(\boldsymbol{s}_j)}{N}-H(S)\right|\geqslant\varepsilon\right\}$$

且 $P(G_\varepsilon)+P(\overline{G}_\varepsilon)=1$。

G_ε 称为 ε 典型序列集，它表示 N 长序列中平均每一维随机变量的自信息非常接近单符号信源熵的一类序列的集合。而$\overline{G}_\varepsilon$ 表示 N 长序列中不在 G_ε 集中的序列的集合，称为非 ε 典型序列集。它们的差别在于$\frac{I(\boldsymbol{s}_j)}{N}$与 $H(\boldsymbol{S})$的差是否小于任意小的正数 ε。

下面推导 ε 典型序列集的一些性质。

（1）G_ε 和$\overline{G}_\varepsilon$ 的概率

$$1\geqslant P(G_\varepsilon)\geqslant1-\delta(N,\varepsilon)$$

$$0\leqslant P(\overline{G}_\varepsilon)\leqslant\delta(N,\varepsilon)$$

（2）G_ε 和$\overline{G}_\varepsilon$ 中序列的概率

根据 ε 典型序列集的定义，G_ε 中序列$\frac{I(\boldsymbol{s}_j)}{N}$与 $H(\boldsymbol{S})$的差小于正数 ε，即

$$-\varepsilon\leqslant\frac{I(\boldsymbol{s}_j)}{N}-H(\boldsymbol{S})\leqslant\varepsilon$$

$$N[H(\boldsymbol{S})-\varepsilon]\leqslant I(\boldsymbol{s}_j)\leqslant N[H(\boldsymbol{S})+\varepsilon]$$

而 $I(\boldsymbol{s}_j)=-\log p(\boldsymbol{s}_j)$，所以

$$2^{-N[H(S)-\varepsilon]}\geqslant p(\boldsymbol{s}_j)\geqslant2^{-N[H(S)+\varepsilon]}$$

（3）G_ε 和$\overline{G}_\varepsilon$ 中序列的个数

设 G_ε 中序列数为 M_G，则

$$1\geqslant p(G_\varepsilon)\geqslant M_G2^{-N[H(S)+\varepsilon]}$$

$$1-\delta(N,\varepsilon)\leqslant P(\overline{G}_\varepsilon)\leqslant M_G2^{-N[H(S)-\varepsilon]}$$

所以

$$[1-\delta(N,\varepsilon)]2^{N[H(S)-\varepsilon]}=\frac{1-\delta(N,\varepsilon)}{2^{-N[H(S)-\varepsilon]}}\leqslant\frac{P(G_\varepsilon)}{2^{-N[H(S)-\varepsilon]}}\leqslant M_G$$

$$\leqslant\frac{P(G_\varepsilon)}{2^{-N[H(S)+\varepsilon]}}\leqslant\frac{1}{2^{-N[H(S)+\varepsilon]}}=2^{N[H(S)+\varepsilon]}$$

即

$$[1-\delta(N,\varepsilon)]2^{N[H(S)-\varepsilon]} \leqslant M_G \leqslant 2^{N[H(S)+\varepsilon]}$$

因此，N 次扩展信源中信源序列可分为两大类，一类是 ε 典型序列，是经常出现的信源序列。当 $N\to\infty$ 时，这类序列出现的概率趋于 1，并且每个 ε 典型序列接近等概分布 $p(\boldsymbol{s}_j) \approx 2^{-N[H(S)]}$。另一类是低概率的非 ε 典型序列，是不经常出现的信源序列。当 $N\to\infty$ 时，这类序列出现的概率趋于 0。

信源的这种划分性质就是渐近等分割性。

G_ε 中的序列虽然是高概率序列，但 G_ε 中的序列数占信源序列总数的比值很小：

$$\xi = \frac{M_G}{q^N} \leqslant \frac{2^{N[H(S)+\varepsilon]}}{q^N} = 2^{-N[\log q - H(S) - \varepsilon]}$$

因为一般 $H(S) < \log q$，所以 $\log q - H(S) - \varepsilon > 0$，$\lim\limits_{N\to\infty} \xi = 0$，信源序列中大部分是不大可能出现的序列。如果我们只对高概率的 ε 典型序列进行一一对应的等长编码，码字总数减少，则所需码长就可减短。

习　题　4

4.1　一信源产生概率为 $p(1)=0.005$，$p(0)=0.995$ 的统计独立二进制数符。这些数符组成长度为 $N=100$ 的数符组。我们对只含有 3 个或少于 3 个“1”的数符组提供一个二进制码字，所有码字的长度 l 相等。

（1）求信源的熵及其冗余度。

（2）求出为所规定的所有符组都提供码字所需的最小码长，比较 $\frac{l}{N}$ 和 $H(S)$。

（3）求信源发出一数符组，而编码器无相应码字的概率。

4.2　设某城市有 805 门公务电话和 60000 门居民电话。作为系统工程师，你需要为这些用户分配电话号码。所有号码均是十进制数，且不考虑电话系统中 0、1 不可用在号码首位的限制。（提示：用即时码概念。）

（1）如果要求所有公务电话号码为 3 位长，所有居民电话号码等长，求居民号码长度 L_1 的最小值。

（2）设城市分为 A、B 两区，其中 A 区有 9000 门电话，B 区有 51000 门电话。现进一步要求 A 区的电话号码比 B 区的短 1 位，求 A 区号码长度 L_2 的最小值。

4.3　设一信源有 6 个符号，概率分别为 $p(s_1)=\frac{1}{2}$，$p(s_2)=\frac{1}{4}$，$p(s_3)=\frac{1}{8}$，$p(s_4)=p(s_5)=\frac{1}{20}$，$p(s_6)=\frac{1}{40}$。求其二元霍夫曼编码和效率。

4.4　设一信源有 3 个符号，概率分别为 0.5、0.4 和 0.1。

（1）求二元霍夫曼编码及效率。

（2）求二次扩展码的霍夫曼编码及效率。

4.5　一信源有 8 个符号，概率分别为 0.2、0.15、0.15、0.1、0.1、0.1、0.1、0.1。求其三进制霍夫曼编码。

4.6 已知5瓶酒中有一瓶变质了（尝起来是苦的）。通过目测观察酒瓶从而判断这些酒坏了的概率分布为$(p_1,p_2,p_3,p_4,p_5)=\left(\frac{1}{3},\frac{1}{4},\frac{1}{6},\frac{1}{6},\frac{1}{12}\right)$。通过品尝可确定哪瓶酒坏了。

（1）假设你每次品尝一瓶，怎样安排合适的品尝顺序，用尽可能少的品尝次数来确定坏酒？并计算品尝次数的均值。

（2）假设你改变了策略，每次不再品尝单独的一瓶酒，而是将数瓶酒混合起来一起品尝，直到找到坏酒为止。首先品尝的应该是哪几瓶酒的混合？答案是唯一的吗？如果是唯一的，请解释为什么；如果不是唯一的，请给出另外一种方案。（提示：品尝数瓶酒的混合等效于品尝其他几瓶酒的混合，因此这不算两种方案。另外，有两瓶酒的坏掉的概率相等，都是$\frac{1}{6}$，所以交换这两瓶酒不算新的方案。）

4.7 设X_1,X_2,X_3为独立的二进制随机变量，并且$P_r\{X_1=1\}=\frac{1}{2}$，$P_r\{X_2=1\}=\frac{1}{3}$，$P_r\{X_3=1\}=\frac{1}{4}$。请给出联合随机变量$X_1X_2X_3$的霍夫曼编码，并求其平均码长。

4.8 下面是4种不同的编码：

（a）$\{000,10,00,11\}$

（b）$\{100,101,0,11\}$

（c）$\{01,100,011,00,111,1010,1011,1101\}$

（d）$\{01,111,011,00,010,110\}$

请分别回答以下问题：

（1）此码的码长分布是否满足Kraft－McMillan不等式？

（2）此码是否是即时码？如果不是，请说明。

（3）此码是否是唯一可译码？如果不是，请说明。

4.9 请问下述编码哪些不可能是一个霍夫曼编码？

（1）$\{0,10,11\}$　　（2）$\{00,01,10,110\}$　　（3）$\{01,10\}$

4.10 即时码在译码时产生的译码延时为0。变长码的译码延时定义为在译码时需要查看的下一个码字的码符号个数的最大值。例如，码$\{0,01\}$的译码延时为1。请给出一个译码延时为3的变长码。

4.11 香农的信源编码定理说明，一个随机变量的紧致码平均码长的下界是它的熵，而在很多情况下实际编码效率要差得多。例如，一个二元随机变量的概率分布是$\{\varepsilon,1-\varepsilon\}$，当$\varepsilon\to0$时，$H(X)=H(\varepsilon)\to0$，而此时它的紧致码平均码长为1。请找出一个随机变量，其熵为2，它的紧致码平均码长3。

4.12 考虑信源分布$\left\{\frac{1}{3},\frac{1}{3},\frac{1}{4},\frac{1}{12}\right\}$。

（1）为此信源构造一个霍夫曼编码。

（2）找出两组不同的最佳编码方案。

（3）用实例说明在最佳码中某些码字的码长会大于它的香农编码的码字长

度 $l_i = \left\lceil \log \frac{1}{p_i} \right\rceil$。

4.13 赛马比赛，共 8 匹马，赢的概率分别为 $\left\{\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{16},\frac{1}{64},\frac{1}{64},\frac{1}{64},\frac{1}{64}\right\}$。如果我们要告诉别人哪匹马会赢，一种方法是给这些马编上号，每匹马要用 3 个二元符号表示。考虑到这些马赢的概率并不相等，有没有更简短的表示方法？

4.14 请找出一个唯一可译码，既不满足前缀条件也不满足后缀条件。

4.15 （1）当 $r=2$ 时，无限长的即时码 $l_1=1$，$l_2=2,\cdots,l_k=k,\cdots$ 是否满足 Kraft 不等式？（2）将上问推广到任意 r 的情况。

4.16 一信源有 4 个符号，概率分布和两种可能的二进制编码如题表 4.16 所示。试回答下列问题（不需计算和证明）：

（1）哪个码是唯一可译码？

（2）哪个码符合即时码条件？

题表 4.16

s_k	$p(s_k)$	1#码	2#码
s_1	0.4	1	1
s_2	0.3	01	10
s_3	0.2	001	100
s_4	0.1	000	1000

4.17 已知信源分布 $\begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 & s_8 \\ \frac{1}{12} & \frac{1}{6} & \frac{1}{12} & \frac{1}{8} & \frac{1}{12} & \frac{1}{4} & \frac{1}{12} & \frac{1}{8} \end{bmatrix}$。

（1）请给出此信源的二进制和三进制费诺编码。

（2）求相应的平均码长、编码效率和码方差。

4.18 现在需要对下面的英文绕口令（共 44 个符号）进行无失真编码：peter piper picked a peck of pickled peppers，其相应的概率分布如题表 4.18 所示。

题表 4.18

符号	**p**	**e**	␣	**c**	**i**	**k**	**r**
次数	9	8	7	3	3	3	3
频率	0.205	0.182	0.159	0.068	0.068	0.068	0.068
符号	**d**	**a**	**f**	**l**	**o**	**s**	**t**
次数	2	1	1	1	1	1	1
频率	0.045	0.023	0.023	0.023	0.023	0.023	0.023

（1）若采用二进制定长编码，请问共需要多少个码符号？编码效率是多少？

（2）理论上对此信源进行编码最少需要多少二进制码符号？

（3）给出此信源的二进制霍夫曼编码。

（4）当使用（3）中的编码方案时，共需多少个码符号？平均码长是多少？编码效率是多少？

4.19 某文本文件包含 4096 个字符，这些字符来自于一个 8 符号的信源符号集合 $S=\{a,b,c,d,e,f,g,h\}$。假设已知此文件中每个字符出现的次数如题表 4.19(1)所示。

题表 4.19(a)

字符	a	b	c	d	e	f	g	h
次数	299	250	501	478	491	512	474	1091

(1) 设计一个二进制霍夫曼编码和一个文件头结构来压缩并传输这个文件，最终需要多少比特？(注：文件的接收者已知信源符号表及压缩策略，但并不知道霍夫曼编码的码表。)

(2) 同样使用(1)中的编码策略，但源文件中各字符出现的次数如题表 4.19(b)所示时，需要多少比特？

题表 4.19(b)

字符	a	b	c	d	e	f	g	h
次数	1024	0	1024	0	1024	0	1024	0

(3) 如果源文件中各字符出现的次数如题表 4.19(c)所示，需要多少比特？

题表 4.19(c)

字符	a	b	c	d	e	f	g	h
次数	4096	0	0	0	0	0	0	0

第 5 章　有噪信道编码

一般信道中总是存在噪声或干扰，信息传输会造成损失，那么有噪信道中能不能无差错的传递信息？如果能够，那么无错误传输的最大信息传输率是多少？这就是本章要研究的内容，即研究通信的可靠性问题。香农在 1948 年的文章中提出并证明了这个极限信息传输率的存在，这个定理叫信道编码定理，也称为香农第二定理。

在有噪信道中信道输入输出之间是统计依赖关系而不是确定关系，因此信道输出要唯一地译成一个输入一般将无法避免差错，发生译码错误的概率称为译码错误概率。对于有噪信道，这个错误概率取决于信道的特性，且不可能为零。但是香农的研究表明，如果把要传输的消息在传输前事先进行编码，并在接收端采用适当的方法译码，则消息有可能得到几乎无误的传输，也就是说，通过不可靠的信道可以实现可靠的信息传输。

5.1　信道编码的相关概念

广义的信道编码指为特定信道传输而进行的传输信号设计与实现，常包括以下几类：

① 描述编码：对特定数据信号的描述，如 NRZ 码、ASCII 码、Gray 码等。

② 约束编码：对于特定信号特性的约束，如用于减少直流分量的 HDB3 码、用于相位与同步检测的 Barker 码等。

③ 扩频编码：用于扩展信号频谱为近似白噪声谱并满足某些相关特性，如 m 序列、Gold 序列等。

④ 纠错编码：用于检测与纠正信号传输过程中因噪声干扰导致的差错，如重复码、循环码、BCH 码、卷积码等，即狭义的信道编码。本书中研究的是狭义的信道编码。

纠错编码又称为差错控制码，虽然与信源编码一样都是一种编码，但信源编码的作用是压缩冗余度以得到信息的有效表示，提高信息传输率，而信道编码的作用是提高信息传输时的抗干扰能力以增加信息传输的可靠性。

在研究信道编码时，信源编码器和信源译码器分别被归于信源和信宿，如图 5.1 所示。

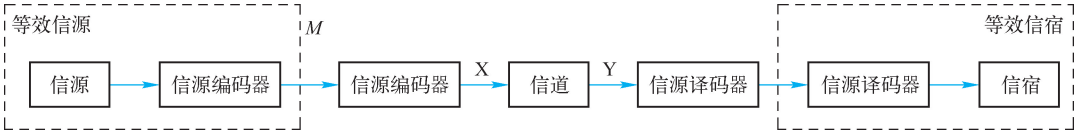


图 5.1　等效通信系统模型

这里假定等效信源送给信道的信源符号是已经经过信源编码的 M 个码字。信道编码的编码对象就是这 M 个信源码字。这 M 个信源码字通常是由二元符号 0、1 构成的码字序列，也叫信息序列，而且经过信源编码后，可以假定符号 0 和 1 是独立等概的。信道编码

就是按一定的规则给信息序列增加一些多余的码元，使不具有规律性的信息序列变为具有某种规律性的信道码字序列 $\boldsymbol{X}$。也就是说，码字序列 $\boldsymbol{X}$ 的码元之间是相关的，在接收端，信道译码器利用这种相关性即已知的编码规则来译码，检验接收到的码字序列 $\boldsymbol{Y}$ 中是否有错，并纠正其中的差错。根据相关性来检测和纠正传输过程中产生的差错就是信道编码的基本思想。

下面分别讨论在有噪信道中信息传输发生错误的概率与什么因素有关系，它们是怎样影响译码错误概率的。

5.1.1 错误概率和译码规则

错误概率与信道的统计特性有关，信道的统计特性可由信道矩阵来表示，由信道矩阵就可以求出错误概率。

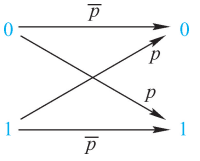


图 5.2　二元对称信道

【例 5.1】在图 5.2 所示的二元对称信道中，单个符号的错误传递概率是 p，单个符号的正确传递概率为 $\overline{p}=1-p$，信道输入的概率分布 $\begin{bmatrix}X\\P\end{bmatrix}=\begin{bmatrix}0 & 1\\ \omega & \overline{\omega}\end{bmatrix}$，可以求出信道输出的概率分布为

$$\begin{cases}P(Y=0)=\omega\,\overline{p}+\overline{\omega}p\\ P(Y=1)=\omega p+\overline{\omega}\,\overline{p}\end{cases}$$

一般，收到 0 后译成 0，收到 1 后译成 1。如果收到 0 实际的信道输入是 1，或者收到 1 后实际的信道输入是 0，则发生了译码错误，因此错误概率为

$$\begin{cases}P(X=1\mid Y=0)=\dfrac{P(X=1,Y=0)}{P(Y=0)}=\dfrac{\overline{\omega}p}{\omega\,\overline{p}+\overline{\omega}p}\\ P(X=0\mid Y=1)=\dfrac{P(X=0,Y=1)}{P(Y=1)}=\dfrac{\omega p}{\omega p+\overline{\omega}\,\overline{p}}\end{cases}$$

平均错误概率为

$$\begin{aligned}P_{\mathrm{E}}&=P(Y=0)P(X=1\mid Y=0)+P(Y=1)P(X=0\mid Y=1)\\&=(\omega\,\overline{p}+\overline{\omega}p)\frac{\overline{\omega}p}{\omega\,\overline{p}+\overline{\omega}p}+(\omega p+\overline{\omega}\,\overline{p})\frac{\omega p}{\omega p+\overline{\omega}\,\overline{p}}\\&=\overline{\omega}p+\omega p\\&=p\end{aligned}$$

因此，错误概率与信道的统计特征有关。

但是通信的过程并不是信息传输到信道输出端就结束了，还要经过译码过程才能到达信宿，译码过程和译码规则对系统的错误概率影响很大。

例如，假定图 5.2 中的二元对称信道 $p=0.9$，其输入符号为等概分布。如果在信道输出端接收到符号 0 时，译码器把它译成 0，接收到 1 时，把它译成 1，那么译码错误概率 $P_{\mathrm{E}}=0.9$。反之，如果规定在信道输出端接收到符号 0 时，译码器把它译成 1，接收到 1 时，把它译成 0，则译码错误概率 $P_{\mathrm{E}}=0.1$。可见，错误概率既与信道统计特性有，也与译码规则有关。

【定义 5-1】设信道的输入符号集 $X=\{x_i\}\ (i=1,2,\cdots,r)$，输出符号集 $Y=\{y_j\}\ (j=1,2,\cdots,s)$，若对每个输出符号 y_j，都有一个确定的函数 $F(y_j)$，使 y_j 对应于唯一的一个输入

符号 x_i，则称这样的函数为译码规则，记为

$$F(y_j)=x_i \qquad i=1,2,\cdots,r;\ j=1,2,\cdots,s$$

对于有 r 个输入、s 个输出的信道而言，输出 y_j 可以对应 r 个输入中的任何一个，所以译码规则共有 r^s 中。

【例 5.2】设有一信道，信道矩阵为

$$\boldsymbol{P}=\begin{bmatrix}0.5 & 0.3 & 0.2\\ 0.2 & 0.3 & 0.5\\ 0.3 & 0.3 & 0.4\end{bmatrix}$$

根据此信道矩阵，可以设计一个译码规则如下

$$A:\begin{cases}F(y_1)=x_1\\ F(y_2)=x_2\\ F(y_3)=x_3\end{cases}$$

也可以设计为另一个译码规则

$$B:\begin{cases}F(y_1)=x_1\\ F(y_2)=x_3\\ F(y_3)=x_2\end{cases}$$

由于 $r=3$，$s=3$，总共可以设计出 $r^s=27$ 种译码规则，应该怎样选择译码规则？

【解】

一个自然的准则就是使平均错误概率为最小。

在确定译码规则 $F(y_j)=x_i$ 后，若信道输出端接收到符号 y_j，则一定译成 x_i，如果发送端发送的确实就是 x_i，就是正确译码；反之，若发送端发送的不是 x_i 就认为是错误译码。于是收到符号 y_j 条件下，译码的正确概率为

$$p[F(y_j)\mid y_j]=p(x_i\mid y_j)$$

而错误概率为

$$\begin{aligned}p(e\mid y_j)&=1-p[F(y_j)\mid y_j]\\&=1-p(x_i\mid y_j)\end{aligned}$$

其中，e 表示除了 $F(y_j)=x_i$ 以外的所有符号的集合。

译码后的平均错误概率 P_{E} 是错误概率 $p(e\mid y_j)$ 对 Y 空间取平均值，即

$$\begin{aligned}P_{\mathrm{E}}&=E[p(e\mid y_j)]\\&=\sum_{j=1}^{s}p(y_j)p(e\mid y_j)\end{aligned}\tag{5.1}$$

它表示经过译码后接收到一个符号平均产生的错误大小。

如何设计译码规则 $F(y_j)=x_i$ 使 P_{E} 最小呢？由于式(5.1)右边是非负项之和，所以选择译码规则使每一项为最小，则所得 P_{E} 为最小。因为 $p(y_j)$ 与译码规则无关，所以只要设计译码规则 $F(y_j)=x_i$ 使错误概率 $p(e\mid y_j)$ 最小，也就是要选择 $p[F(y_j)\mid y_j]$ 最大。这就是最大后验概率准则。

【定义 5-2】选择译码函数 $F(y_j)=x^*$，使之满足条件

$$p(x^*\mid y_j)\geqslant p(x_i\mid y_j)\quad(\forall i,x^*\in X)\tag{5.2}$$

称为最大后验概率译码规则，又称为最小错误概率准则、最优译码、最佳译码。它是对于每个输出符号 $y_j(j=1,2,\cdots,s)$ 均译成具有最大后验概率的那个输入符号 x^*，这样译码平均错误概率最小。

一般是已知信道的前向概率 $p(y_j \mid x_i)$ 和输入符号的先验概率 $p(x_i)$，但后验概率不知，所以最大后验概率译码规则使用起来不是很方便。

根据贝叶斯定律，式(5.2)又可写成

$$\frac{p(y_j \mid x^*)p(x^*)}{p(y_j)} \geqslant \frac{p(y_j \mid x_i)p(x_i)}{p(y_j)} \quad (\forall i)$$

一般，$p(y_j) \neq 0$。这样，最大后验概率译码规则就可表示为：选择译码函数 $F(y_j)=x^*$，使满足 $p(y_j \mid x^*)p(x^*) \geqslant p(y_j \mid x_i)p(x_i)$，$x_i \in X$，即

$$p(x^* y_j) \geqslant p(x_i y_j) \tag{5.3}$$

当输入符号的先验概率 $p(x_i)$ 相等时，式(5.3)又可写成

$$p(y_j \mid x^*) \geqslant p(y_j \mid x_i)$$

因此我们又定义了一个极大似然译码规则。

【定义 5-3】选择译码函数 $F(y_j)=x^*$ 使 $p(y_j \mid x^*) \geqslant p(y_j \mid x_i)$ $(\forall i, x^* \in X)$，称为极大似然译码规则。

根据最大似然译码准则我们可以直接从信道矩阵的转移概率中去选定译码函数。收到 y_j 后，译成信道矩阵 $\boldsymbol{P}$ 第 j 列中最大的转移概率所对应的 x_i。

当输入符号等概时，这两个译码规则是等价的，均可以使平均错误概率 P_E 最小。如果先验概率不相等或不知道时，采用极大似然译码规则不一定能使 P_E 最小。

根据上述译码规则，下面来推导计算平均错误概率的多种表达式：

$$\begin{aligned} P_E &= \sum_j p(e \mid y_j)p(y_j) \\ &= \sum_j \{1 - p[F(y_j) \mid y_j]\}p(y_j) \\ &= \sum_j \sum_{i \neq *} p(y_j \mid x_i)p(x_i) \\ &= \sum_{Y, X-x^*} p(xy) \end{aligned} \tag{5.4}$$

共 $(r-1)s$ 项求和。求和号下面的 $X-x^*$ 表示在输入符号集 X 中对 x^* 以外的所有元素求和。式(5.4)表示对联合概率矩阵中除 $p(x^* y_j)(j=1,2,\cdots,s)$ 以外的所有元素求和。

平均正确概率为

$$\begin{aligned} \overline{P}_E &= 1 - P_E \\ &= \sum_j p[F(y_j) y_j] \\ &= \sum_j p(x^* y_j) \end{aligned} \tag{5.5}$$

式(5.4)又可写成

$$P_E = \sum_{Y, X-x^*} p(y_j \mid x_i)p(x_i)$$

如果输入等概，即 $p(x_i)=\frac{1}{r}$，则

$$P_E = \frac{1}{r}\sum_{Y,X-x^*} p(y_j \mid x_i) \tag{5.6}$$

式(5.6)表明，在输入等概的情况下，译码错误概率可用信道矩阵中的元素 $p(y_j \mid x_i)$ 求和来表示，在除去每列中对应于 $F(y_j)=x^*$ 那一项后，求矩阵中其余元素之和。

【例 5.3】讨论当输入为等概分布和不等概分布两种情况下，例 5.2 中两种译码规则对应的平均错误概率。

【解】

译码规则 B 就是极大似然译码规则。在输入为等概分布时，极大似然译码规则可使平均错误概率最小。

当输入为等概分布时，两种译码规则所对应的平均错误概率分别为

$$\begin{aligned} P_E(A) &= \frac{1}{3}\sum_{Y,X-x^*} p(y_j \mid x_i) \\ &= \frac{1}{3}[(0.2+0.3)+(0.3+0.3)+(0.2+0.5)] = 0.6 \\ P_E(B) &= \frac{1}{3}\sum_{Y,X-x^*} p(y_j \mid x_i) \\ &= \frac{1}{3}[(0.2+0.3)+(0.3+0.3)+(0.2+0.4)] = 0.567 \end{aligned}$$

当输入不为等概分布时，假设某个输入概率分布 $p(x_1)=\frac{1}{4}$，$p(x_2)=\frac{1}{4}$，$p(x_3)=\frac{1}{2}$，则

$$P'_E(A)=\frac{1}{4}(0.3+0.2)+\frac{1}{4}(0.2+0.5)+\frac{1}{2}(0.3+0.3)=0.6$$

$$P'_E(B)=\frac{1}{4}(0.3+0.2)+\frac{1}{4}(0.2+0.3)+\frac{1}{2}(0.3+0.4)=0.6$$

当输入不为等概分布时，最大似然译码准则的平均错误概率不是最小。最小错误概率译码准则可以得到最小的平均译码错误概率。

根据式(5.3)，由其联合概率矩阵

$$\boldsymbol{P}=\begin{bmatrix} 0.125 & 0.075 & 0.05 \\ 0.05 & 0.075 & 0.125 \\ 0.15 & 0.15 & 0.2 \end{bmatrix}$$

可得，译码函数

$$C:\begin{cases} F(y_1)=x_3 \\ F(y_2)=x_3 \\ F(y_3)=x_3 \end{cases}$$

此时的平均错误概率为

$$P_E(C)=(0.125+0.05)+(0.075+0.075)+(0.05+0.125)=0.5$$

发生译码错误是由于信道中的噪声，信道噪声的影响使得在接收端收到输出符号 Y 后对发送端发送的符号 X 仍然存在不确定性，因此平均错误概率与信道疑义度存在着一定的关系，这个关系用费诺不等式表示。

【定理 5-1】平均错误概率 P_E 与信道疑义度 $H(X \mid Y)$ 满足以下关系：

$$H(X \mid Y) \leqslant H(P_E) + P_E \log(r-1)$$

【证明】因为

$$P_E = \sum_{Y, X-x^*} p(xy)$$

$$\overline{P}_E = 1 - P_E = \sum_{Y} p(x^* y)$$

所以

$$\begin{aligned} H(P_E) + P_E \log(r-1) &= P_E \log \frac{1}{P_E} + (1-P_E) \log \frac{1}{1-P_E} + P_E \log(r-1) \\ &= P_E \log \frac{r-1}{P_E} + (1-P_E) \log \frac{1}{1-P_E} \\ &= \sum_{Y, X-x^*} p(xy) \log \frac{r-1}{P_E} + \sum_{Y} p(x^* y) \log \frac{1}{1-P_E} \end{aligned}$$

而信道疑义度

$$\begin{aligned} H(X \mid Y) &= \sum_{XY} p(xy) \log \frac{1}{p(x \mid y)} \\ &= \sum_{Y, X-x^*} p(xy) \log \frac{1}{p(x \mid y)} + \sum_{Y} p(x^* y) \log \frac{1}{p(x^* \mid y)} \end{aligned}$$

所以

$$\begin{aligned} & H(X \mid Y) - H(P_E) - P_E \log(r-1) \\ &= \sum_{Y, X-x^*} p(xy) \log \frac{P_E}{(r-1) p(x \mid y)} + \sum_{Y} p(x^* y) \log \frac{1-P_E}{p(x^* \mid y)} \end{aligned}$$

因为 $\log x \leqslant x-1$，所以

$$\begin{aligned} H(X \mid Y) - H(P_E) - P_E \log(r-1) &\leqslant \left\{ \sum_{Y, X-x^*} p(xy) \left[\frac{P_E}{(r-1) p(x \mid y)} - 1 \right] + \sum_{Y} p(x^* y) \left[\frac{1-P_E}{p(x^* \mid y)} - 1 \right] \right\} \cdot \log_2 e \\ &= \left\{ \frac{P_E}{r-1} \sum_{Y, X-x^*} p(y) - \sum_{Y, X-x^*} p(xy) + (1-P_E) \sum_{Y} p(y) - \sum_{Y} p(x^* y) \right\} \cdot \log_2 e \\ &= \left\{ \frac{P_E}{r-1} \sum_{X-x^*} \sum_{Y} p(y) - P_E + (1-P_E) - (1-P_E) \right\} \cdot \log_2 e \\ &= \left\{ \frac{P_E}{r-1} (r-1) - P_E \right\} \cdot \log_2 e \\ &= 0 \end{aligned}$$

证毕。

虽然 P_E 与译码规则有关，但是不管采用什么译码规则该不等式都是成立的。费诺不等式表明，接收到 Y 后关于 X 的平均不确定性可以分为两部分，第一部分 $H(P_E)$ 是指接收到 Y 后是否产生错误的不确定性，第二部分 $P_E \log(r-1)$ 是当错误 P_E 发生后，判断是那个输入符号造成错误的最大不确定性，是 $r-1$ 个符号不确定性的最大值与 P_E 的乘积。若以 P_E 为横坐标，$H(P_E) + P_E \log(r-1)$ 是随 P_E 变化的曲线，如图 5.3 所示，$H(X \mid Y)$ 的值在曲线的下方。

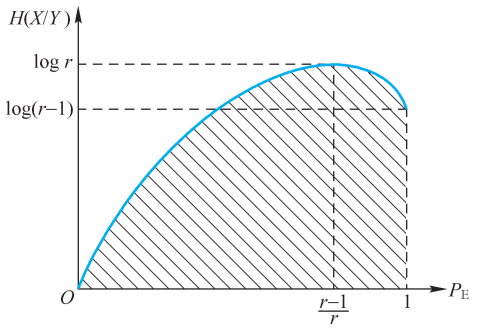


图 5.3　费诺不等式曲线图

P_E 的最大值为 1，这时 $H(X|Y) \leqslant \log(r-1)$，当 $P_E = \dfrac{r-1}{r}$时，曲线取到最大值，$H(X|Y) \leqslant \log r$。

5.1.2　错误概率与编码方法

前面讨论了平均错误概率 P_E 与译码规则的关系，选择最佳译码规则可以减小错误概率 P_E。下面讨论通过择恰当的编码方法可以进一步减小错误概率 P_E。

1. 简单重复编码

设有二元对称信道（见图 5.2），相应的信道矩阵为

$$\boldsymbol{P} = \begin{bmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{bmatrix}$$

选择最佳译码规则为

$$\begin{cases} F(y_1) = x_1 \\ F(y_2) = x_2 \end{cases}$$

总的平均错误概率在输入分布为等概的条件下为

$$P_E = \frac{1}{r} \sum_{Y, X=x^*} p(y|x)$$

$$= \frac{1}{2}(0.01 + 0.01) = 10^{-2}$$

对于一般数字通信系统，该错误概率是非常大的，数字通信一般要求错误概率在 $10^{-9} \sim 10^{-6}$的范围内，有的甚至要求更低的错误概率。那么，在上述统计特性的二元信道中，是否有办法使错误概率降低呢？实际经验告诉我们：只要在发送端把消息重复发几遍，就可使接收端接收消息时错误减小，从而提高通信的可靠性。

例如，发信源符号 0 时，可以重复发送 3 个 0，发 1 时，重复发送 3 个 1，这可以看成离散无记忆信道的三次扩展信道。这样在信道输入端有两个码字 000 和 111，在输出端由于信道干扰，各码元都可能发生错误，则有 8 个可能的输出序列。输入是在三次扩展信道的 8 个二元序列中选两个作为消息，而输出端这 8 个二元序列都是可能的接收序列。这时信道矩阵为

$$\boldsymbol{P}=\begin{array}{c}\\ 000\\ 111\end{array}\begin{array}{c}\begin{array}{cccccccc}000 & 001 & 010 & 011 & 100 & 101 & 110 & 111\end{array}\\ \begin{bmatrix}\bar{p}^3 & \bar{p}^2p & \bar{p}^2p & \bar{p}p^2 & \bar{p}^2p & \bar{p}p^2 & \bar{p}p^2 & p^3\\ p^3 & \bar{p}p^2 & \bar{p}p^2 & \bar{p}^2p & \bar{p}p^2 & \bar{p}^2p & \bar{p}^2p & \bar{p}^3\end{bmatrix}\end{array}$$

假设输入符号为等概分布，采用极大似然译码规则。如果 p 远小于 1，则接收序列与译成的发送码字的对应关系如表 5.1 所示。

表 5.1　接收序列与译成的发送码字的对应

接收序列	译码输出	接收序列	译码输出
000	000	011	111
001		101	
010		110	
100		111	

译码后的平均错误概率（$p=0.01$）：

$$\begin{aligned}P_{\mathrm{E}}&=\frac{1}{2}\left[p^3+\bar{p}p^2+\bar{p}p^2+\bar{p}p^2+3\,\bar{p}p^2+\bar{p}p^2+\bar{p}p^2+p^3\right]\\&=p^3+3\,\bar{p}p^2\approx 3\times10^{-4}\end{aligned}$$

这个译码函数也可以直观地根据输出端接收序列是 0 多还是 1 多来判断，如果有两个以上是 0，则译码器判决为 0；如果有两个以上是 1，则判决为 1，即“择多译码”。

根据择多译码规则，同样可得到

$$\begin{aligned}P_{\mathrm{E}}&=\text{错 3 个码元的概率}+\text{错 2 个码元的概率}\\&=C_3^3p^3+C_3^2\,\bar{p}p^2\\&=p^3+3\,\bar{p}p^2\approx3\times10^{-4}\end{aligned}$$

与原来的二元对称信道的平均错误概率 10^{-2} 相比，这种简单重复编码（重复三次）的平均错误概率降低了近两个数量级。这是因为若接收码字中有一位码元发生错误，译码器还能正确译出所发送的码字，若传输中 2 位或 3 位码元发生错误时，译码器才会译错。所以，这种简单重复编码能纠正 1 位码元的错误，使得错误概率降低。

显然，如果进一步增大重复次数 n，则会继续降低平均错误概率。可算得

$$\begin{aligned}&n=5 && P_{\mathrm{E}}=10^{-5}\\&n=7 && P_{\mathrm{E}}=4\times10^{-7}\\&n=9 && P_{\mathrm{E}}=10^{-8}\\&n=11 && P_{\mathrm{E}}=5\times10^{-10}\end{aligned}$$

可见，当 n 很大时，使 P_{E} 很小是可能的。但这带来了一个新问题，当 n 很大时，信息传输会降低很多。我们把经过信道编码后的信息传输率表示为

$$R=\frac{\log M}{n}\text{比特/码符号}$$

这是因为一般假定 M 个信源符号（序列）已接近等概分布，则平均每个信源符号（序列）所携带的信息量为 $\log M$ 比特，用 n 个码元的信道编码码字来传输，平均每个码符号所携带的信息量即为信息传输率 R。

如果传输每个码符号平均需要 t 秒，则信道编码后的信息传输速率为

$$R_t = \frac{\log M}{nt} \text{比特/秒}$$

当 $M=2$ 时，可依次求得简单重复编码的信息传输率：

$$n=1 \qquad R=1$$

$$n=3 \qquad R=\frac{1}{3}$$

$$n=5 \qquad R=\frac{1}{5}$$

$$\vdots \qquad \vdots$$

$$n=11 \qquad R=\frac{1}{11}$$

由此可见，利用简单重复编码减小平均错误概率 P_E 是以降低信息传输率 R 为代价的，那么，怎样编码可以使平均错误概率 P_E 充分小而信息传输率又不致太小呢？

2. (5,2)线性码

我们首先看一下简单重复编码为什么使信息传输率降低。在未重复以前，输入端有 2 个消息，$M=2$。假设为等概率分布，则每个消息携带的信息量是 $\log M=1$ 比特。

$n=3$ 的简单重复编码后，可以把信道看成是无记忆信道的三次扩展信道，这时输入端有 8 个二元序列可以作为消息，但是我们只选择了 2 个二元序列作为消息，$M=2$，每个消息携带的平均信息量仍为 1 比特，而传送一个消息需要 3 个二元码符号，所以 R 就降低到$\frac{1}{3}$比特/码符号。

如果在扩展信道的输入端把 8 个二元序列都用上，则 $M=8$，每个消息平均携带的信息量就是 $\log M=\log 8=3$ 比特，而传递一个消息仍需 3 个二元码，这样 R 就提高到 1 比特/码符号。译码时，接收端 8 个接收序列译成与它对应的发送序列，只要接收序列中有一个码元发生错误，就会变成其他码字序列，使译码造成错误。只有接收序列中每个码元都不发生错误，才能正确传递，所以得到正确传递的概率为$\bar{p}^3$。于是错误概率为

$$P_E = 1-\bar{p}^3 = 3\times 10^{-2} \quad (p=0.01)$$

这时的 P_E 反而比单符号信道传输的 P_E 大 3 倍。

因此，在一个二元信道的 n 次无记忆扩展信道中，输入端有 2^n 个符号序列可以作为消息。如果选出其中的 M 个作为消息传递，则当 M 大一些，R 就大些，P_E 也大些。M 取小一些，R 就要降低，P_E 也降低。对于简单重复编码来说，这似乎是个不可调和的矛盾。

若在三次无记忆扩展信道中，取 $M=4$，用如下 4 个符号序列作为消息：000，011，101，110。信息传输率为

$$R = \frac{\log 4}{3} = \frac{2}{3} \text{比特/符号}$$

按照最大似然译码规则，可计算出错误概率为 $P_E=2\times 10^{-2}$。

与 $M=8$ 的情况相比，错误概率降低了，而信息传输率也降低了。再进一步看，从 2^3 个符号序列中取 $M=4$ 个作为消息，可以有 C_8^4 共 70 种选择方法。

现在来比较 $M=4$ 的第二种选法：000，001，010，100。第二种选法的错误概率为

$$P_E = 2.28 \times 10^{-2}$$

$$R = \frac{2}{3} \text{比特/符号}$$

选取的方法不同，则编码方法不同，错误概率不同。对于第一种码，当发送码字中任一码元发生错误，译码时就可判断消息在传输中发生了错误，但无法判断由哪个消息发生错误而来。对于第二种码，当发送码字 000 时，传输中任一码元发生错误就变成了其他 3 个可能的发送码字，根本无法判断传输时有无发生错误。可见，错误概率与编码方法有很大关系。

例如，信道输入端所选的消息数不变，即取 $M=4$，增加码字的长度，取 $n=5$。这时信道为二元对称信道的五次扩展信道，在信道输入端 $2^5=32$ 个二元序列中选取其中 4 个作为发送码字。这时信息传输率为

$$R = \frac{\log 4}{5} = \frac{2}{5} \text{比特/符号}$$

设输入序列 $\boldsymbol{x}_i = x_{i_1}x_{i_2}x_{i_3}x_{i_4}x_{i_5}$，$x_{i_k} \in \{0,1\}$ $(i=1,2,3,4)$，其中 x_{i_k} 为 $\boldsymbol{x}_i$ 序列中第 k 个分量。若 $\boldsymbol{x}_i$ 中各分量满足方程

$$\begin{cases} x_{i_1} = x_{i_1} \\ x_{i_2} = x_{i_2} \\ x_{i_3} = x_{i_1} \oplus x_{i_2} \\ x_{i_4} = x_{i_1} \\ x_{i_5} = x_{i_1} \oplus x_{i_2} \end{cases}$$

其中，$\oplus$ 为模二和运算，也叫异或。写成矩阵形式为

$$\begin{bmatrix} x_{i_1} \\ x_{i_2} \\ x_{i_3} \\ x_{i_4} \\ x_{i_5} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_{i_1} \\ x_{i_2} \\ x_{i_3} \\ x_{i_4} \\ x_{i_5} \end{bmatrix}$$

由上述编码方法得到一种(5,2)线性码：00000，01101，10111，11010。如果译码采用最大似然译码规则，它的译码规则如表 5.2 所示。

表 5.2　(5,2)线性码译码规则

接收码字	译码输出	接收码字	译码输出
00000	00000	10111	10111
00001		10110	
00010		10101	
00100		10011	
01000		11111	
10000		00111	
10001		00110	
00011		10100	

续表

接收码字	译码输出	接收码字	译码输出
01101	01101	11010	11010
01100		11011	
01111		11000	
01001		11110	
00101		10010	
11101		01010	
11100		01011	
01110		11001	

这种编码方法，接收端译码时能纠正码字中所有发生一位码元的错误，也能纠正其中二个二位码元的错误，所以正确译码概率为

$$\overline{P}_{\mathrm{E}}=\overline{p}^{5}+5\,\overline{p}^{4}p+2\,\overline{p}^{3}p^{2}$$

错误译码概率为

$$\begin{aligned}P_{\mathrm{E}}&=1-\overline{P}_{\mathrm{E}}\\&=1-\overline{p}^{5}-5\,\overline{p}^{4}p-2\,\overline{p}^{3}p^{2}\\&=7.8\times10^{-4}\,(p=0.01)\end{aligned}$$

与前述 $M=4$、$n=3$ 的两种编码方法相比，这种编码方法虽然信息传输率略低，但错误概率减少很多。再与 $M=2$、$n=3$ 的简单重复码相比较，它们的错误概率接近同一个数量级，但(5,2)线性码的信息传输率却比 $n=3$ 的简单重复码的信息传输率高。因此，增大 n，并且适当增大 M 并采用恰当的编码方法，既能使 P_{E} 降低，又能使信息传输率不至于太低。

下面先引入码字距离的概念，再解释(5,2)线性码能获得较低 P_{E} 的原因。

3. 汉明距离

【定义 5-4】长度为 n 的两个符号序列（码字）$\boldsymbol{x}_i$ 与 $\boldsymbol{y}_j$ 之间的距离是指序列 $\boldsymbol{x}_i$ 和 $\boldsymbol{y}_j$ 对应位置上码符号不同的个数，通常又称为汉明距离，用 $d(\boldsymbol{x}_i,\boldsymbol{y}_j)$ 表示。

例如，二元序列 $\boldsymbol{x}_i=101111$，$\boldsymbol{y}_j=111100$，则 $D(\boldsymbol{x}_i,\boldsymbol{y}_j)=3$；四元序列 $\boldsymbol{x}_i=1320120$，$\boldsymbol{y}_j=1220310$，则 $d(\boldsymbol{x}_i,\boldsymbol{y}_j)=3$。

对于如下二元码序列

$$\begin{aligned}\boldsymbol{x}_i&=x_{i_1}x_{i_2}\cdots x_{i_n}\qquad x_{i_k}\in\{0,1\}\\\boldsymbol{y}_j&=y_{j_1}y_{j_2}\cdots y_{j_n}\qquad y_{j_k}\in\{0,1\}\end{aligned}$$

则 $\boldsymbol{x}_i$ 和 $\boldsymbol{y}_j$ 的汉明距离可表示为

$$d(\boldsymbol{x}_i,\boldsymbol{y}_j)=\sum_{k=1}^{n}x_{i_k}\oplus y_{j_k}$$

码字之间的距离越大，则由一个码字变成为另一个码字的可能性越小。当码间距离为 1 时，表示它们在逻辑空间中是相邻的。对于一个码长为 n 的码字，它有 n 个相邻的码字。

这样定义的码字距离满足距离公理，即汉明距离满足以下性质：

① 非负性：$d(\boldsymbol{x}_i,\boldsymbol{y}_j)\geqslant0$，当且仅当 $\boldsymbol{x}_i=\boldsymbol{y}_j$ 时等号成立。

② 对称性：$d(\boldsymbol{x}_i,\boldsymbol{y}_j)=d(\boldsymbol{y}_j,\boldsymbol{x}_i)$。

③ 三角不等式：$D(\boldsymbol{x}_i,\boldsymbol{z}_k)+D(\boldsymbol{y}_j,\boldsymbol{z}_k)\geqslant D(\boldsymbol{x}_i,\boldsymbol{y}_j)$。

【定义 5-5】在码 C 中，任意两个码字的汉明距离的最小值称为该码的最小距离，即

$$d_{\min}=\min\{d(w_i,w_j)\}\ (w_i\neq w_j,w_i,w_j\in C)$$

码的最小距离 $d_{\min}$ 与译码错误概率有关。

我们用距离概念来考察以下 5 个码，如表 5.3 所示。显然，$d_{\min}$ 越大，P_E 越小。码的最小距离 $d_{\min}$ 越大，受干扰后，越不容易把一个码字错成另一个码字，因此错误概率小；反之，若 $d_{\min}$ 越小，受干扰后越容易把一个码字变成另一个码字，因此错误概率大。这告诉我们：在编码选择码字时，要使码字之间的距离越大越好。

表 5.3　码的最小距离与平均译码错误概率

	码 1	码 2	码 3	码 4	码 5
码　　字	000 111	000 011 101 110	000 001 010 100	00000 01101 10111 11010	000 001 010 011 100 101 110 111
消息数 M	2	4	4	4	8
信息传输率 R	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{2}{5}$	1
码的最小距离 $d_{\min}$	3	2	1	3	1
错误概率 P_E （最大似然译码）	3×10^{-4}	2×10^{-2}	2.28×10^{-2}	7.8×10^{-4}	3×10^{-2}

现在还可以把汉明距离与最大似然译码规则联系起来，用汉明距离来表述极大似然译码准则。

极大似然译码准则是对于 $\forall i$，选择译码函数 $F(\boldsymbol{y}_j)=\boldsymbol{x}^*$，使 $p(\boldsymbol{y}_j\mid\boldsymbol{x}^*)\geqslant p(\boldsymbol{y}_j\mid\boldsymbol{x}_i)$。设码字 $\boldsymbol{x}_i=x_{i_1}x_{i_2}\cdots x_{i_n}$，$\boldsymbol{y}_j=y_{j_1}y_{j\,2}\cdots y_{j_n}$，在传输过程中发送码字 $\boldsymbol{x}_i$ 中有 d_{ij} 个位置发生错误，接收端接收序列为 $\boldsymbol{y}_j$，即 $d(\boldsymbol{x}_i,\boldsymbol{y}_j)=d_{ij}$，没有发生错误的位置有 $n-d_{ij}$ 个。

当二元对称信道是无记忆时，有

$$\begin{aligned}p(\boldsymbol{y}_j\mid\boldsymbol{x}_i)&=p(y_{j_1}\mid x_{i_1})p(y_{j_2}\mid x_{i_2})\cdots p(y_{j_n}\mid x_{i_n})\\&=p^{d_{ij}}\cdot\overline{p}^{(n-d_{ij})}\end{aligned}$$

只要 $p<\frac{1}{2}$（正常情况，如 $p=10^{-2}$），则 d_{ij} 越大，$p(\boldsymbol{y}_j\mid\boldsymbol{x}_i)$ 越小；d_{ij} 越小，$p(\boldsymbol{y}_j\mid\boldsymbol{x}_i)$ 越大。

因此，二元对称信道中极大似然译码规则可用汉明距离表示为：选择译码函数 $F(\boldsymbol{y}_j)=\boldsymbol{x}^*$，使 $d(\boldsymbol{x}^*,\boldsymbol{y}_j)\leqslant d(\boldsymbol{x}_i,\boldsymbol{y}_j)$，即

$$d(\boldsymbol{x}^*,\boldsymbol{y}_j)=\min_i(\boldsymbol{x}_i,\boldsymbol{y}_j)$$

也就是在接收到码字 $\boldsymbol{y}_j$ 后，在输入码字集中 $\{\boldsymbol{x}_i\}\ (i=1,2,\cdots,r)$ 中寻找一个与 $\boldsymbol{y}_j$ 的汉明距离最小的码字 $\boldsymbol{x}^*$，这又称为最小距离译码规则。前面提到的择多译码就是一种最小距离译码的方法。

这时，平均译码错误概率也可用汉明距离来表示。设输入码字数为 M（并设输入等概

分布），则

$$P_{\mathrm{E}} = \frac{1}{M}\sum_{\boldsymbol{Y},\boldsymbol{X}-\boldsymbol{x}^*} p(\boldsymbol{y}\mid\boldsymbol{x}) = \frac{1}{M}\sum_{j}\sum_{i\neq *} p^{d_{ij}}(1-p)^{n-d_{ij}}$$

或者

$$P_{\mathrm{E}} = 1-\frac{1}{M}\sum_{\mathbf{Y}} p(\boldsymbol{y}\mid\boldsymbol{x}^*) = 1-\frac{1}{M}\sum_{j} p^{d_{*j}}(1-p)^{n-d_{*j}}$$

其中，$d_{*j}=d(\boldsymbol{x}^*,\boldsymbol{y}_j)$。

在非二元对称信道中可采用最小距离译码规则，但它不一定等价于极大似然译码规则。

当输入为等概时，由于极大似然译码规则与最大后验概率译码规则是等价的，所以这时最小汉明距离译码规则与最大后验概率译码规则也是等价的。

从上面的讨论可知，在 M 和 n 相同的情况下，即保持一定的信息传输率 R 时，选择不同的编码方法码的最小距离也不同，我们选择码的最小距离最大的那一个码。在译码时，则将接收序列译成与其距离最小的码字，这样得到的 P_{E} 最小。那么只要码长 n 足够长，总可以通过恰当的选择 M 个码字使 P_{E} 很小，而 R 保持一定的水平。

5.2 有噪信道编码定理

【定理 5-2】设有一个离散无记忆平稳信道，其信道容量为 C。当信息传输率 $R<C$ 时，只要码长 n 足够长，则总存在一种编码，可以使译码错误概率 P_{E} 任意小。否则，如果 $R>C$，则无论 n 取多大，也找不到一种编码，使译码错误概率 P_{E} 任意小。

这个定理称为有噪信道编码定理，又称为香农第二定理。有噪信道编码定理告诉我们，如果我们编码码长为 n，选用的码字个数 $M\leqslant 2^{n(C-\varepsilon)}$，$\varepsilon$ 为任意小的正数，则编码后，信道的信息传输率为

$$R=\frac{\log M}{n}\text{比特/码符号}$$

$R<C$，所以可以在有噪声干扰的信道中，以任意小的错误概率传输信息，而且当 n 足够大时，可以以任意接近信道容量 C 的信息传输率传输信息。若选用码字总数 $M\geqslant 2^{n(C+\varepsilon)}$，则无论 n 取多大，也找不到一种编码，使译码错误概率 P_{E} 任意小。

通过一个有噪信道可以实现几乎无失真传输，这与人们的直观想象似乎大相径庭，而这个定理的证明也是非常巧妙的。按照通常的思路证明这一结论可能先要构造一个理想的好码，然后计算这个码用于传输时的平均错误概率，但这两点都难以实现。首先，构造具有理想性能的好码是一个非常复杂的问题，在当时的计算条件下无法解决。其次，想在 n 很大时计算这一理想好码在最佳译码或极大似然译码规则下的 P_{E} 也是极其困难的。香农巧妙地避开了这两个难题，首先，他不去构造理想的好码，而是用随机编码的方法得到所有可能生成的码的集合，然后在码集合中随机选择一个码作为输入码序列，最后计算这样随机选择的一个码在码集合上的平均性能。这样算出的 P_{E} 可以达到任意小，因而可以证明一定存在一种

编码，它的性能 P_E 达到或者超过随机编码的平均性能。证明的基本思路如下：

① 允许平均错误概率任意小而非零。

② 在 n 次无记忆扩展信道中讨论，n 足够大，这样可以使用大数定理。

③ 随机编码，在随机编码的基础上计算整个码集的码的平均错误概率，由此证明至少有一种好码存在。因为是随机编码，所以求错误译码概率时与特定的码字无关。

所谓随机编码，是指在 n 长的输入序列中，随机选择 M 个作为输入码字组成一个码 $C=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_M\}$，$M$ 为信源消息数。每次选择一个码字有 2^n 种可能的选择，共 M 个码字，所以共有 $(2^n)^M$ 种可能的码，也就是通过随机编码可以得到 2^{nM} 个码。这是一个很大的数，如 $M=2^8$、$n=16$ 时，$2^{nM}=2^{4096}\approx 10^{1233}$。当然，在这些码中有一部分是无法用的，比如某些码的码字有重复，但由于码字个数为 $M=2^{nR}$，这只占全部可能的码字序列 2^n 中的很小的一部分，因此同一码中出现相同码字的概率很小，我们可以忽略这个问题。

由于所求的是平均性能，就可以用大数定律且不必考虑码的结构。在译码时，将接收序列译成与其联合典型的码字。这种译码方法不是最优译码，但便于理论分析。

以上定理只是在离散无记忆信道的情况下证明的，但是对连续信道和有记忆信道结论同样成立。

香农第二定理也只是一个存在定理，它说明错误概率趋于零的好码是存在的，但是没有说明如何构造这个好码。尽管如此，香农第二定理仍然具有重要的理论意义和实践指导作用，可以指导各种通信系统的设计，有助于评价各种通信系统及编码效率。

从香农第一定理和香农第二定理可以看出，要做到有效和可靠的传递信息，需要进行信源编码和信道编码。首先，通过信源编码，用尽可能少的信道符号来表达信源，也就是对信源数据用最有效的表达方式表达，尽可能减少编码后数据的冗余度。然后，对信源编码后的数据设计信道编码，即适当增加一些冗余度，以纠正和克服信道中干扰引起的错误。这两部分是分别独立考虑的。

分两部分编码的方法在实际通信系统中有着重要的意义。现代通信系统大多数都是数字通信系统，比模拟通信系统有许多优点。在实际数字通信系统中，信道常常是共用的数字信道（二元信道），语音、音乐、图像、数据都用同一通信信道来传输。因此，可以将语音、图像先数字化，再对数字化的语音、图像等信源进行不同的信源编码，针对各自信源的不同特点，用不同的数据压缩方法。对于共同的数字信道来说，输入端只是二元序列，所以信道编码只需针对不同的信道特性进行，以纠正信道传输带来的错误，这样可以大大降低通信系统设计的复杂度。

离散无记忆信道中 P_E 趋于零的速度与 n 成指数关系。当 $R<C$ 时，平均错误概率为

$$P_E\leqslant\exp[-nE_r(R)]$$

其中，$E_r(R)$ 称为随机编码指数，又称为可靠性函数或加拉格（Gallagher）函数。一般，可靠性函数 $E_r(R)$ 与信息传输率 R 的关系曲线如图 5.4 所示，它是一条下凸函数曲线。在 $R<C$ 范围内，$E_r(R)>0$，所以随 n 增大 P_E 以指数趋于零。实际编码的码长 n 不需选得很大，P_E 就能很快趋于零。

可靠性函数 $E_r(R)$ 在信道编码中有极其重要的意义，表示在码长 n 已定时，P_E 的上界。在实际问题中，为了达到一定的可靠性，即要求 P_E 小于某个值（如 10^{-6}），可靠性函数 $E_r(R)$ 可以帮助我们选择信息传输率和编码长度 n。

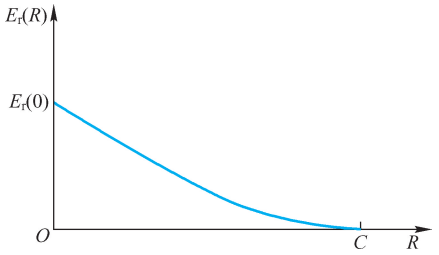


图 5.4　$E_r(R)$ 曲线图

综合上述定理和论述可知，信道的信道容量是可靠传输的分界点：当 $R<C$ 时，P_E 以指数趋于零，当 $R>C$ 时，P_E 很快趋于 1。因此，在任何信道中，C 是可靠传输的最大的信息传输率。

5.3　纠错编码

香农的信道编码定理引起了人们对信道编码的极大兴趣，但是香农只是证明了满足这种特性（$R<C$ 时，$P_E \to 0$）的码的存在，还不能按其证明方法得到这种好码。证明过程是采用随机编码的方法，由于随机编码所得的码集很大，通过搜索得到好码的方法在实际上很难实现；而且即使找到其中的好码，这种码的码字也是毫无结构的，这意味着译码时只能用查表的方法，而在 N 很大时，译码表所需的存储量也是很难被接受的，所以真正实用的信道编码还需通过各种数学工具来构造，使码具有很好的结构性，以便译码。抽象代数（也称为近世代数）就是编码理论的最重要的数学工具，包括群论、环论、域论、格论、线性代数等许多分支。

信道编码的目的是提高信号传输的可靠性，广义的信道编码还包括为特定信道设计的传输信号，如 NRZ（不归零）码、HDB3 码、伪随机序列码都属于信道编码，而纠错编码作为提高传输可靠性的最主要措施之一，是我们研究的主要内容。

5.3.1　纠错编码分类

由于信道中干扰和噪声存在，使得经信道传输后的接收码字与原来的发送码字存在差异，也就是差错。一般，信道中噪声干扰越大，码字产生错误的概率也越大。

信道中的干扰一般分为两种形式：一是随机噪声，主要来源于设备的热噪声和散弹噪声以及传播媒介的热噪声，它是通信系统中的主要噪声；二是脉冲干扰和信道衰落，其特点是突发出现，主要来源于雷电、通电开关、负荷突变或设备故障等。

根据干扰和噪声形式，信道可分为三类：随机信道、突发信道和混合信道。

随机噪声产生的错误是独立随机出现的，称为随机错误。其特点是各码元是否发生错误是随机的，且相互独立，因此不会出现成片的错误。只产生随机错误的信道称为随机信道。这是比较典型的常见信道。以高斯白噪声为主体的信道属于这类信道，如卫星信道、同轴电缆、光缆信道和大多数微波中继信道。

脉冲干扰和信道衰落产生的错误是成串出现的，产生的错误之间有相关性，这类错误被

称为突发错误。产生突发错误的信道称为突发信道。实际的短波信道、移动通信信道、由于擦伤造成成串差错的光盘和磁盘，均为突发信道。

有些实际信道既有随机错误又有突发错误，称为混合信道。

对不同类型的信道要设计不同类型的信道编码才能收到良好效果。根据不同的信道类型设计的信道编码分为纠独立随机错误码、纠突发错误码和纠混合错误码。在通信系统中，纠检错的工作方式有反馈重传纠错、前向纠错和混合纠错。

1. 反馈重传纠错

反馈重传纠错方式如图 5.5 所示。发送端发出的是能够发现错误的检错码，接收端对接收到的信息进行译码，发现有错时，通过反馈系统向发送端请求重传已发送的全部或部分码字，直到接收端认为无错误为止。我国的电报系统就是一种反馈重传纠错系统。

2. 前向纠错

前向纠错也称为自动纠错，发送端发出的是具有纠错能力的纠错码，接收端根据编码规则进行解码（如图 5.6 所示）。当误码个数在码的纠错能力范围内，译码器可以自动纠正错误。根据信道的不同类型，设计不同的纠错编码。

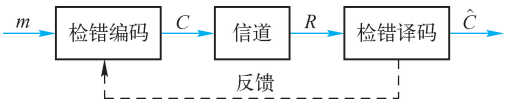


图 5.5　反馈重传纠错

图 5.6　前向纠错

3. 混合纠错

对发送端进行适当编码，当错误不严重时，在码的纠错能力之内，采用自动纠错，如果超出了码的纠错能力，则通过反馈系统向发送端要求重发，同时具有反馈重传纠错和自动纠错工作方式的纠错称为混合纠错。

检错码和纠错码在不加区别时统称为纠错码。纠错编码的目的是引入剩余度，就是在传输的信息码元后增加一些多余的码元（称为校验元），以使信息损失或发生传输错误后仍然能在接收端恢复。信息序列用 $\boldsymbol{m}$ 表示，纠错编码输出序列为 $\boldsymbol{C}$，通过信道传输后接收码序列为 $\boldsymbol{R}$，通过纠错以后我们希望恢复的码序列$\hat{\boldsymbol{C}}=\boldsymbol{C}$。

从理论上讲，编码实现的是信息序列到纠错码序列的映射。具体实现时，由于考虑时延的限制和计算复杂度的限制，只能将信息序列分组后按一定的映射关系变换成输出码序列。

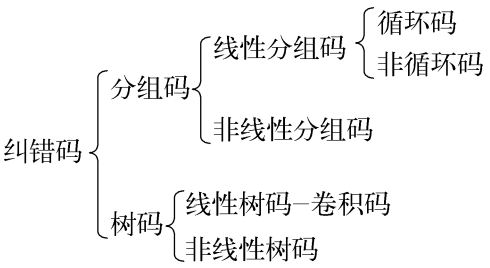
根据不同的分组方式及其随后的映射关系，纠错码可以分为分组码和树码。

分组码：把信息序列以每 k 个码元分组，然后把每组 k 个信息元按一定规律产生 r 个多余的校验元，输出序列每组长为 $n=k+r$。每个码字的 r 个校验元只与本组的 k 个信息元有关，与别的分组的信息位无关，记为分组码(n,k)。

树码：信息序列以每 k_0（通常较小）个码元分段，编码器输出该段的校验元不仅与本段的 k_0 个信息元有关，而且与其前面若干段的信息元有关，称为树码或链码。树码中最重要的一类是校验元与信息元的关系是线性关系，称为卷积码。

根据信息码元与校验码元之间是否存在线性关系，纠错码可以分为线性码和非线性码。线性码的校验码元是若干位信息码元的线性组合，非线性码的校验位与信息位不满足线性关系。线性码具有良好的数学结构，编译码比较简单，性能优于同样纠错能力的非线性码。

本章主要介绍线性分组码。线性分组码中有一类特别重要的是循环码，它具有完整的代数结构，编译码都比较简单和易于实现。纠错码分类结构如下：



因为目前的通信系统大多是二进制的数字系统，所以不特别说明，以下提到的纠错码都是指二元码。

5.3.2 纠错编码的基本概念

通常，信源编码把信源符号用二元序列来表示，这个二元序列称为信息序列。信源编码主要解决的是通信的有效性问题，即用尽可能少的码符号来表示信源符号或信源符号序列。信息序列送到信道传输以前，还需要经过信道编码，变成具有纠检错能力的码序列。信道编码要解决的问题是通信的可靠性的问题。通过在信息序列中插入冗余码元（称为校验元或监督元），使新序列的码元之间具有相关特性，然后进行传输。在接收端，信道译码器根据这个相关特性对接收序列进行译码，在纠错能力范围内，可以对差错进行自动纠正，恢复原发送码序列。

对信道编码的一般要求如下。

① 纠错检错能力强，可发现和纠正多个错误。

② 信息传输率高。信息传输率也称为码率，$R=\dfrac{\log M}{n}$，M 为信息序列的个数，所以码长 n 应尽可能短。信息传输率表示每个码元符号所携带的信息量。

③ 编码规律简单，实现设备简单且费用合理。

④ 与信道的差错统计特性相匹配。

信道编码就是在综合考虑以上因素的情况下，选择和设计合理的编译码实现方案。

将信源编码器的输出序列进行分组，分组长度为 k，则可以有 $M=2^k$ 个不同的分组信息序列。每个分组信息序列用一个 n 长的码字来表示（$n>k$），$\boldsymbol{C}=[C_1 \quad C_2 \quad \cdots \quad C_n]$，这样的 2^k 个码字的集合称为二元分组码。

每个码字 $\boldsymbol{C}=[C_1 \quad C_2 \quad \cdots \quad C_n]$ 中 k 位称为信息位，其余 $n-k$ 位为校验位或监督位。例如，当 $k=3$，$n=7$ 时，最大的信息序列数 $M=2^3=8$ 个，而长为 $n=7$ 的二元码共有 $2^7=128$ 个，选出其中的 8 个作为码字，也称为允用序列。其他序列为禁用序列。

通常，用 $\eta=\dfrac{k}{n}$ 表示码字中信息位所占的比重，称为编码效率或码率。η 越大，编码效率越高，它是衡量码性能的一个重要参数。

【例 5.4】下面给出一种编码：

消息序列	码字
000	0000000
001	0011101
010	0100111
011	0111010
100	1001110
101	1010011
110	1101001
111	1110100

信息位 $k=3$，码长 $n=7$，监督位 $r=4$，$\eta=\frac{3}{7}=43\%$。而 $R=\frac{\log M}{n}=\frac{\log 2^k}{n}=\frac{k}{n}=\frac{3}{7}$，恰好等于编码效率。这是因为这个码充分利用了 3 位的信息位生成了 8 个码字。

如果 n 长码字的每一位与原始信息序列的 k 个信息位之间的函数关系是线性关系，则称该分组码为线性分组码，否则称为非线性分组码。

若(n,k)分组码中 k 个信息位与原始信息序列的 k 个信息位相同，且位于 n 长码字的前（或后）k 位，而校验位位于其后（或前）$n-k$ 位，则该分组码为系统码，否则为非系统码。

对于某一码字，其非零元素的个数称为该码字的汉明重量。

对于二元码，其码字的重量是码字中 1 的个数。码字 $\boldsymbol{c}_i=[c_{i_1}\quad c_{i_2}\quad \cdots\quad c_{i_n}]$ 的重量可以表示为 $W(\boldsymbol{c}_i)=\sum_{k=1}^{n} c_{i_k}$。例如，码字 $\boldsymbol{c}_1=1010101$，其重量为 4。

一个线性分组码中的任意两个码字的和仍然是它的一个码字。因此，码字距离与重量之间的关系为

$$\begin{aligned} d(\boldsymbol{c}_i,\boldsymbol{c}_j) &= W(\boldsymbol{c}_i \oplus \boldsymbol{c}_j) \\ &= \sum_{k=1}^{n} c_{i_k} \oplus c_{j_k} \end{aligned}$$

5.3.3 线性分组码

线性分组码是最有实用价值的一类码，如汉明码（Hamming Code）、高莱码（Golay Code）、RS 码（Reed－Solomon code）、BCH 码等。线性分组码的编码方式是将信源输出序列分组，每组是长为 k 的信息序列，然后按照一定的编码规则插入 $n-k$ 位的校验位，校验位是所有信息位的线性组合，组成 n 长的码字序列。

1. 校验矩阵与生成矩阵

线性分组码可以由一组信息元的模 2 线性方程生成。例如，一个$(7,3)$线性分组码 $\boldsymbol{C}=[C_1\quad C_2\quad C_3\quad C_4\quad C_5\quad C_6\quad C_7]$，其中 C_1、C_2、C_3 为信息元，C_4、C_5、C_6、C_7 为校验元。假设校验元可用下面方程组得到

$$\begin{cases} C_4 = C_1 + C_3 \\ C_5 = C_1 + C_2 + C_3 \\ C_6 = C_1 + C_2 \\ C_7 = C_2 + C_3 \end{cases}$$

⊕即模 2 加。这是一组线性方程，确定了由信息元得到校验元的规则，所以称为校验方程或监督方程。

方程组还可以写成矩阵形式。首先，将方程组改写，使方程的等号右边为 0。

$$\begin{cases} C_1 + C_3 + C_4 = 0 \\ C_1 + C_2 + C_3 + C_5 = 0 \\ C_1 + C_2 + C_6 = 0 \\ C_2 + C_3 + C_7 = 0 \end{cases}$$

然后，写成矩阵相乘的形式

$$\begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \\ C_3 \\ C_4 \\ C_5 \\ C_6 \\ C_7 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{5.7}$$

令

$$\boldsymbol{H} = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

则式(5.7)可写成

$$\boldsymbol{H}\boldsymbol{C}^{\mathrm{T}} = \boldsymbol{0} \qquad 或 \qquad \boldsymbol{C}\boldsymbol{H}^{\mathrm{T}} = \boldsymbol{0}$$

其中，$\boldsymbol{H}$ 称为一致校验矩阵。

一旦建立了校验矩阵，校验元与信息元的关系就确定了，码字也随之确定。

校验方程还可以改写成：

$$\begin{cases} C_1 = C_1 \\ C_2 = C_2 \\ C_3 = C_3 \\ C_4 = C_1 + C_3 \\ C_5 = C_1 + C_2 + C_3 \\ C_6 = C_1 + C_2 \\ C_7 = C_2 + C_3 \end{cases} \tag{5.8}$$

令 $\boldsymbol{m}=[C_1 \quad C_2 \quad C_3]$，则

$$\boldsymbol{G}=\begin{bmatrix}1&0&0&1&1&1&0\\0&1&0&0&1&1&1\\0&0&1&1&1&0&1\end{bmatrix}$$

式(5.8)可写成

$$\boldsymbol{C}=\boldsymbol{mG}$$

其中，$\boldsymbol{C}$ 为 n 维行向量，$\boldsymbol{m}$ 为 k 维行向量。$\boldsymbol{G}$ 为 $k\times n$ 矩阵，称为线性分组码 $\boldsymbol{C}$ 的生成矩阵。利用生成矩阵可将信息序列 $\boldsymbol{m}$ 变成码字序列 $\boldsymbol{C}$。例如，当 $\boldsymbol{m}=[0 \quad 1 \quad 1]$时，可以得到

$$\begin{aligned}\boldsymbol{C}&=\boldsymbol{mG}\\&=[0\quad 1\quad 1]\begin{bmatrix}1&0&0&1&1&1&0\\0&1&0&0&1&1&1\\0&0&1&1&1&0&1\end{bmatrix}\\&=[0\quad 1\quad 1\quad 1\quad 0\quad 1\quad 0]\end{aligned}$$

为方便起见，可以将生成矩阵写成 $\boldsymbol{G}=\begin{bmatrix}\boldsymbol{G}_1\\\boldsymbol{G}_2\\\vdots\\\boldsymbol{G}_k\end{bmatrix}$的形式，其中 $\boldsymbol{G}_i(i=1,2,\cdots,k)$ 为 n 维行向量 $\boldsymbol{G}_i=[g_{i_1} \quad g_{i_2} \quad \cdots \quad g_{i_n}]$。

对于信息序列 $\boldsymbol{m}=[m_1 \quad m_2 \quad \cdots \quad m_k]$，有

$$\boldsymbol{C}=\boldsymbol{mG}=\sum_{i=1}^{k}m_i\boldsymbol{G}_i \tag{5.9}$$

式(5.9)表明，码字 $\boldsymbol{C}$ 为信息序列 $\boldsymbol{m}$ 和生成矩阵 $\boldsymbol{G}$ 的行向量的线性组合，这里的和为“模2加”。当信息组 $\boldsymbol{m}$ 中只有一个非零元素时，码字为生成矩阵的某一行，因此生成矩阵的每一行都是一个码字，k 个不相同的码字可以构成码的生成矩阵，由这 k 个码字的不同线性组合便生成了整个码组。所选取的 k 个码字必须是线性无关的，也就是生成矩阵的秩为 k，才能由生成矩阵组合出 2^k种不同的码字，这样生成的全部码字组成了 n 维矢量空间的一个 k 维子空间。矩阵 $\boldsymbol{G}$ 的 k 个行向量是这一子空间的 k 个基矢量，而 k 维子空间的任意 k 个线性无关的矢量都可以作为这一子空间的基矢量。这些基矢量组成的矩阵都可以看成这一子空间或这一码的生成矩阵，同一码的所有可能的生成矩阵都是等价的。

$\boldsymbol{G}$ 可以写成分块矩阵，即 $\boldsymbol{G}=[\boldsymbol{I} \quad \boldsymbol{P}]$。$\boldsymbol{I}$ 为 $k\times k$ 的单位矩阵，$\boldsymbol{P}$ 为 $k\times(n-k)$的一般矩阵。这样生成的码 $\boldsymbol{C}$ 是系统码，信息位在码字的前 k 位。当生成矩阵不能写成由 k 阶单位矩阵构成的分块矩阵时，生成的码 $\boldsymbol{C}$ 不是系统码。但根据矩阵理论，可以将一般形式的矩阵通过行初等变成系统形式的矩阵，两个矩阵是等价的，因此这样产生的码与系统码是等价的，每一个线性码对应唯一一个系统形式的生成矩阵。所以，以系统码为研究对象不失一般性。

由于系统形式的生成矩阵有 k 阶单位子阵，因此组成系统形式的生成矩阵的 k 个码字是线性无关的。化成系统形式之后，容易验证生成矩阵的各行是否线性无关。

由于生成矩阵 $\boldsymbol{G}$ 的每一行都是一个码字，所以生成矩阵和校验矩阵有如下关系：

$$\boldsymbol{HG}^{\mathrm{T}}=\boldsymbol{0}^{\mathrm{T}} \quad \text{或} \quad \boldsymbol{GH}^{\mathrm{T}}=\boldsymbol{0}$$

即线性分组码的生成矩阵和校验矩阵的行矢量彼此正交。以上结果表明，线性分组码既可以由生成矩阵确定，也可以由校验矩阵确定。(n,k)线性分组码是 n 维 n 长向量构成的线性空间中一个 k 维线性子空间，它可以由 $\boldsymbol{G}$ 或 $\boldsymbol{H}$ 确定。同时，$\boldsymbol{H}$ 矩阵的行矢量在 n 维矢量空间中张成一个 $n-k$ 维子空间。这两个子空间的矢量是互相垂直的。$n-k$ 维子空间也对应一个线性码，这一码与 $\boldsymbol{G}$ 生成的码互为对偶码。$\boldsymbol{G}$ 为 $n-k$ 维子空间的校验矩阵。

一般要构造一个(n,k)线性分组码，只要找出一个秩为 $n-k$ 的 $n-k$ 行、n 列矩阵 $\boldsymbol{H}$，则可由齐次线性方程组 $\boldsymbol{HC}^{\mathrm{T}}=\boldsymbol{0}^{\mathrm{T}}$ 的解空间的全部向量作为许用码字，得到一个(n,k)线性分组码。因此，线性分组码可以用齐次线性方程组这样方便的数学工具来研究。

标准形式的校验矩阵可以写成 $\boldsymbol{H}=[\boldsymbol{Q}\quad \boldsymbol{I}]$ 形式，$\boldsymbol{Q}$ 为$(n-k)\times k$ 矩阵，$\boldsymbol{I}$ 为$(n-k)\times(n-k)$单位矩阵。而系统形式的生成矩阵可以写成 $\boldsymbol{G}=[\boldsymbol{I}\quad \boldsymbol{P}]$，所以

$$\boldsymbol{HG}^{\mathrm{T}}=[\boldsymbol{Q}\quad \boldsymbol{I}]\begin{bmatrix}\boldsymbol{I}\\ \boldsymbol{P}^{\mathrm{T}}\end{bmatrix}=\boldsymbol{Q}+\boldsymbol{P}^{\mathrm{T}}=\boldsymbol{0}^{\mathrm{T}}$$

可得

$$\boldsymbol{P}^{\mathrm{T}}=\boldsymbol{Q}\qquad 或\qquad \boldsymbol{P}=\boldsymbol{Q}^{\mathrm{T}}$$

因此，系统形式的校验矩阵和生成矩阵可以很方便地实现转换。

线性分组码的性质如下：

① 码中任意两个码字之和仍为一码字。

② 任意码字是 $\boldsymbol{G}$ 的行向量 $\boldsymbol{G}_1,\boldsymbol{G}_2,\cdots,\boldsymbol{G}_k$ 的线性组合。

③ 零向量 $\boldsymbol{0}=[0\ 0\ \cdots\ 0]$ 是一个码字，称为零码字。

④ 线性分组码的最小距离等于非零码字的最小重量。

线性分组码最重要的性质是其线性特性以及在此基础上的对称性。所谓线性特性，是指线性码中任意两个码字的和或差仍为一码字。对称性是指在一个码的所有码字上减去一个特定的码字，结果仍是同一码的全部码字。这样在求码字间的距离分布时，只需求出任一码字与其他所有码字的距离分布即可。

【例 5.5】重复码是一个(3,1)线性分组码。其生成矩阵为

$$\boldsymbol{G}=[1\quad 1\quad 1]$$

$$\boldsymbol{C}=C_1C_2C_3=[m_1][1\quad 1\quad 1]=[m_1\quad m_1\quad m_1]$$

【例 5.6】偶校验码是一个(4,3)线性分组码，其生成矩阵为

$$\boldsymbol{G}=\begin{bmatrix}1&0&0&1\\0&1&0&1\\0&0&1&1\end{bmatrix}$$

$$\begin{aligned}\boldsymbol{C}&=[C_1\ C_2\ C_3\ C_4]\\&=[m_1\quad m_2\quad m_3]\begin{bmatrix}1&0&0&1\\0&1&0&1\\0&0&1&1\end{bmatrix}\\&=[m_1\quad m_2\quad m_3\quad m_1+m_2+m_3]\end{aligned}$$

【例 5.7】已知生成矩阵为

$$\boldsymbol{G}=\begin{bmatrix}1&0&0&1&1&1&0\\0&1&0&0&1&1&1\\0&0&1&1&1&0&1\end{bmatrix}$$

求生成的线性分组码及由 $\boldsymbol{H}$ 生成的线性分组码。

【解】

由于 $\boldsymbol{G}=[\boldsymbol{I}\quad\boldsymbol{P}]$，有

$$\boldsymbol{P}=\begin{bmatrix}1&1&1&0\\0&1&1&1\\1&1&0&1\end{bmatrix}$$

又因为 $\boldsymbol{Q}=\boldsymbol{P}^{\mathrm{T}}$，则

$$\boldsymbol{H}=[\boldsymbol{Q}\quad\boldsymbol{I}]=\begin{bmatrix}1&0&1&1&0&0&0\\1&1&1&0&1&0&0\\1&1&0&0&0&1&0\\0&1&1&0&0&0&1\end{bmatrix}$$

由生成矩阵 $\boldsymbol{G}$ 生成的(7,3)线性分组码为：

$\boldsymbol{m}$	$\boldsymbol{C}$
000	0000000
001	0011101
010	0100111
011	0111010
100	1001110
101	1010011
110	1101001
111	1110100

把校验矩阵 $\boldsymbol{H}$ 当作生成矩阵，可生成(7,4)线性分组码：

$\boldsymbol{m}$	$\boldsymbol{C}'$
0000	0000000
0001	0110001
0010	1100010
0011	1010011
0100	1110100
0101	1000101
0110	0010110
0111	0100111
1000	1011000
1001	1101001
1010	0111010
1011	0001011
1100	0101100
1101	0011101
1110	1001110
1111	1111111

这样生成的 $\boldsymbol{C}$ 和 $\boldsymbol{C}'$是互相正交的。

2. 线性分组码的纠检错能力

由生成矩阵产生的码字在信道传输过程中，由于干扰的存在，使得一些码元发生错误。接收端通过编码规则进行译码，如能发现错误，则称为检错，如果再能纠正错误，称为纠错。码能纠、检错误码元的个数称为该码的纠检错能力。如果发现错误和纠正错误的个数越多，则说明该码的纠、检错能里力强。只要接收码字没有错到变成其他发送码字，就可以发现错误。如果不会错判成其他发送码字，就可以纠正错误正确译码。因此，我们设计的码字之间应有较大的区别，即它们的汉明距离要大。

线性分组码的最小距离等于非零码字的最小重量。因为根据线性分组码的封闭性可知，任意两个码字的和仍是一个码字。根据码字之间的距离的定义，两个码字和的非零个数即为它们之间的距离，而两个码字和的非零个数又是新码字的重量。所以，线性分组码的最小距离必为它的非零码字的最小重量。

码的最小距离越大，即码中任意两个码字之间的差别越大，越不容易把一个码字误传成其他码字，译码时也容易正确译码，因此码的纠、检错能力越大，即码的最小距离和非零码字的最小重量决定了码的纠检错能力。

关于码的最小距离与纠、检错能力的关系有以下结论：

【定理 5-3】对于(n,k)线性分组码，设 $d_{\min}$为码的最小汉明距离，则

（1）这组码有纠正 u 个错误的能力的充要条件是 $d_{\min}=2u+1$。

（2）能检测 l 个错误的充要条件是 $d_{\min}=l+1$。

（3）能纠正 u 个错误，同时可以发现 l（$l>u$）个错误的充要条件为 $d_{\min}=u+l+1$。

【证明】

（1）这组码有纠正 u 个错误的能力，也就是只要发生小于等于 u 个错误，译码的时候就能正确地译成原来的码字，而不会错纠成其他码字。

如图 5.7(a)所示，如果分组码中任意两个码字之间的汉明距离 $d(\boldsymbol{x}_i,\boldsymbol{x}_j)\geqslant d_{\min}=2u+1$，那么码字发生小于等于 u 个错误的时候，仍然在它的纠错范围内（可以用半径为 u 的球体表示码字的纠错范围），而不会进入另一个码字的纠错范围内，所以 $d_{\min}=2u+1$ 的线性分组码具有纠正 u 个错误的能力。反过来说，如果这组码有纠正 u 个错误的能力，那么需要码字之间的距离至少需要 $2u+1$，两个球体之间的距离至少为 1，才能保证发生小于等于 u 个错误的时候，不会进入另一个码字的纠错范围内。

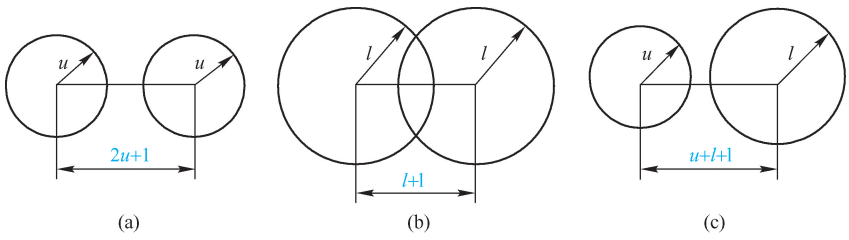


图 5.7　码的最小距离与纠、检错能力

（2）能检测 l 个错误，就意味着一个码字发生小于等于 l 个错误能够被发现，而不会被认为是其他码字。这就要求以一个码字为球心，以 l 为半径的球与其他码字的距离至少为 1，

这样发生小于等于 l 个错误的时候，才不会变成其他码字，所以可以被检测到。否则，会变成其他码字，就不能被检测到了，如图 5.7（b）所示。

（3）这是一种纠检结合的工作方式。若发生小于等于 u 个错误则按前向纠错方式工作，以节省反馈重发的时间。发生大 u 个错误，则超过了该码的纠错能力，自动按检错重发方式工作，如图 5.7（c）所示。

对于任意码字 $d(\boldsymbol{x}_i,\boldsymbol{x}_j)\geqslant d_{\min}=u+l+1$，$l>u$，由前面的讨论可知，该码发生小于等于 u 个错误，显然满足（1）的条件，所以具有纠正小于等于 u 个错误的能力，而只要错误码元的个数小于等于 l 就不会落入其他码字的纠错范围，不会被错纠，因此可以发现大于等于 l 个错误。所以 $d_{\min}=u+l+1$ 的线性分组码能纠正 u 个错误，同时可以发现 $l(l>u)$ 个错误。

例如，$d_{\min}=5$，按检错方式工作时，$l=4$；按纠错方式工作时 $u=2$；按纠检结合方式工作时，$u=1$，$l=3$，这时，当发送码发生 1 个错码时可纠，发生 2 个或 3 个错误时可检，发生 4 个错码时就会落入另一个发送码的纠错范围内，被错纠了。

例如，表 5.4 所示的简单重复码(3，1)线性分组码，$k=1$，$l=3$。$d_{\min}=3$ 可纠一个错误。

表 5.4　(3,1)简单重复码

m	C	R
0	000	000
		001
		010
		100
1	111	011
		101
		110
		111

定理 5-3 说明了码的最小距离与纠检错能力的关系。选择编码方案时应选择极大最小距离码可以得到较高的纠检错能力。

3. 校验矩阵与最小距离的关系

码的最小距离或者说码的纠错能力与它的校验矩阵列向量的线性相关程度有关。

【定理 5-4】对于(n,k)线性分组码，若校验矩阵 $\boldsymbol{H}$ 中的任意 t 列线性无关而$t+1$列线性相关，则码的最小距离或码字的最小重量为 $t+1$；若码字的最小重量或码的最小距离为 $t+1$则 $\boldsymbol{H}$ 的任意 t 列线性无关而 $t+1$ 列线性相关。

【证明】

把 $\boldsymbol{H}$ 写成行向量的形式 $\boldsymbol{H}=[\boldsymbol{H}_1\quad \boldsymbol{H}_2\quad \cdots\quad \boldsymbol{H}_n]$，其中 $\boldsymbol{H}_j$ 为列向量。

对于任意的码字 $\boldsymbol{C}$，有以下关系：

$$\begin{aligned}\boldsymbol{H}\boldsymbol{C}^{\mathrm{T}}&=[\boldsymbol{H}_1\ \boldsymbol{H}_2\cdots\boldsymbol{H}_n][C_1\ C_2\cdots C_n]^{\mathrm{T}}\\&=C_1\boldsymbol{H}_1+C_2\boldsymbol{H}_2+\cdots+C_n\boldsymbol{H}_n\\&=\boldsymbol{0}\end{aligned}$$

$\boldsymbol{H}\boldsymbol{C}^{\mathrm{T}}$ 等于 n 个 $n-k$ 维列向量 $\boldsymbol{H}_j$ 的线性组合，并且由 $\boldsymbol{H}\boldsymbol{C}^{\mathrm{T}}=\boldsymbol{0}$ 可知，n 个向量 $\boldsymbol{H}_j$ 是线性相关的。

若 $\boldsymbol{H}$ 中任意 t 列线性无关而 $t+1$ 列线性相关，则说明 $C_1,C_2,\cdots,C_n$ 这 n 个码符号中必有 $t+1$ 个 1，其余为 0，使得 $t+1$ 个向量 $\boldsymbol{H}_j$ 的线性组合等于 $\boldsymbol{0}$，即 $\boldsymbol{C}$ 的码重为 $t+1$。而且，$\boldsymbol{C}$ 的码重不可能为 t，否则可使 t 个向量 $\boldsymbol{H}_j$ 的线性组合等于 $\boldsymbol{0}$，即 $\boldsymbol{H}$ 中有 t 列线性相关，这与前面的假定矛盾。

反之，如果码字的最小重量为 $t+1$，则重量最小的码字有 $t+1$ 个非零码元，代入 $\boldsymbol{H}\boldsymbol{C}^{\mathrm{T}}=\boldsymbol{0}$，

则有 $t+1$ 个 $\boldsymbol{H}_j$ 线性组合等于 $\mathbf{0}$。这说明必有 $t+1$ 个向量 $\boldsymbol{H}_j$ 线性相关，并且任意 t 个向量必定线性无关。因为如果有 t 个 $\boldsymbol{H}$ 的列向量线性相关，则必然存在一个重量为 t 的码字。这与最小码重为 $t+1$ 是矛盾的。

证毕。

由于 $\boldsymbol{H}$ 是 $(n-k)\times n$ 矩阵，其秩至多为 $n-k$，即最多有 $n-k$ 个列向量线性无关。在寻找好码时，我们希望 $d_{\min}$ 越大越好，就希望 $\boldsymbol{H}$ 中线性无关的列向量越多越好。而线性无关的列向量最多为 $n-k$ 个，所以 $d_{\min}\leqslant n-k+1$。

如果设计的 (n,k) 线性分组达到了 $d_{\min}=n-k+1$，则称为极大最小距离码。

4. 线性分组码的伴随式及伴随式译码

根据校验方程 $\boldsymbol{C}\boldsymbol{H}^{\mathrm{T}}=\mathbf{0}$，校验矩阵可以用来验证接收码字是否为我们的许用码字。

设发送码字为 $\boldsymbol{C}$，接收码字为 $\boldsymbol{R}$。令 $\boldsymbol{S}=\boldsymbol{R}\boldsymbol{H}^{\mathrm{T}}$。当 $\boldsymbol{R}$ 为许用码字时因满足校验方程，所以 $\boldsymbol{S}=\boldsymbol{R}\boldsymbol{H}^{\mathrm{T}}=0$。若 $\boldsymbol{S}\neq\mathbf{0}$，则说明 $\boldsymbol{R}$ 不是一个发送码字，码字在传输过程中发生了错误。因此 $\boldsymbol{S}$ 是码字在传输过程中是否出现错误的标志，称为伴随式（或称为监督子、校验子等）。

因为接收码字 $\boldsymbol{R}$ 是由发送码字 $\boldsymbol{C}$ 在传输过程中产生差错得到的，所以可以将 $\boldsymbol{R}$ 写成 $\boldsymbol{R}=\boldsymbol{C}+\boldsymbol{E}$。$\boldsymbol{E}=[e_1\ e_2\cdots e_n]$ 称为差错图样。当码字的第 i 位发生错误时 $e_i=1$，否则 $e_i=0$。这样，伴随式又可以写成

$$\boldsymbol{S}=[\boldsymbol{C}+\boldsymbol{E}]\boldsymbol{H}^{\mathrm{T}}=\boldsymbol{C}\boldsymbol{H}^{\mathrm{T}}+\boldsymbol{E}\boldsymbol{H}^{\mathrm{T}}=\boldsymbol{E}\boldsymbol{H}^{\mathrm{T}}$$

可以看出，伴随式仅与错误图样有关，与码字无关，即伴随式中仅包含错误图样信息。$\boldsymbol{S}=\mathbf{0}$，则表示传输中要么无差错发生，要么错误图案恰好为一个码字，而差错图案恰为一个码字的机会是很小的。如果 $\boldsymbol{S}\neq 0$，可以由伴随式得到错误图样信息，然后对接收码字加以修正，以得到正确译码。

伴随式 $\boldsymbol{S}$ 是伴随接收码的一个 $n-k$ 维向量，但是从 $\boldsymbol{S}=\boldsymbol{E}\boldsymbol{H}^{\mathrm{T}}$ 可以看出，$\boldsymbol{S}$ 并不反映发送的码字是什么，而只是反映信道对码字造成“怎样的干扰”。差错图样 $\boldsymbol{E}$ 是 n 维矢量，有 2^n 个可能的组合，而伴随式 $\boldsymbol{S}$ 是 $n-k$ 维矢量，只有 2^{n-k} 个可能的组合，因此不同的差错图样可能有相同的伴随式。

在接收端，我们并不知道发码 $\boldsymbol{C}$ 是什么，但可以知道 $\boldsymbol{H}^{\mathrm{T}}$ 和 $\boldsymbol{R}$ 是什么，并通过伴随式译码找到 $\boldsymbol{C}$ 的估值。其过程是 $\boldsymbol{S}=\boldsymbol{R}\boldsymbol{H}^{\mathrm{T}}=\boldsymbol{E}\boldsymbol{H}^{\mathrm{T}}\Rightarrow\boldsymbol{E}\Rightarrow\boldsymbol{C}=\boldsymbol{R}+\boldsymbol{E}$，即先算出 $\boldsymbol{S}$，再由 $\boldsymbol{S}$ 算出 $\boldsymbol{E}$。最后令 $\boldsymbol{C}=\boldsymbol{R}+\boldsymbol{E}$ 求出 $\boldsymbol{C}$。关键是如何从 $\boldsymbol{S}$ 找出 $\boldsymbol{E}$。

因为 $\boldsymbol{S}^{\mathrm{T}}=\boldsymbol{H}\boldsymbol{E}^{\mathrm{T}}=\sum_{i=1}^{n}e_iH_i$，$\boldsymbol{H}_i$ 为 $\boldsymbol{H}$ 的列向量，所以伴随式是接收码字中发生错误的码元在 $\boldsymbol{H}$ 中对应列的矢量和。由于同一个伴随式会对应多个差错图样，根据最小距离译码规则，求到 $\boldsymbol{S}$ 后应该译成其中重量最小的错误图样。伴随式有 2^{n-k} 个，其中 1 个对应没有差错的图样，n 个对应 n 个码元中发生了一个错误的图样，C_n^2 个对应 n 个码元中发生了两个差错的图样，C_n^3 个对应 n 个码元中发生了 3 个差错的图样，C_n^u 个对应 n 个码元中发生了 u 个差错的图样，直到把 2^{n-k} 个伴随式用尽，所以

$$2^{n-k}\geqslant\sum_{i=0}^{u}C_n^i$$

因为假设该线性分组码可以纠正 u 个错误码元。

把 $\boldsymbol{S}$ 和 $\boldsymbol{E}$ 以及 $\boldsymbol{R}$、$\boldsymbol{C}$ 的对应关系列成一个表，称为标准阵列（如表 5.5 所示）。通过查标准阵列译码，可以很快得到发送码字 $\boldsymbol{C}$。下面举例说明这个译码过程。

表 5.5　标准阵列译码

$\boldsymbol{S}$	$\boldsymbol{R}$ 与 $\boldsymbol{E}$ 对 $\boldsymbol{C}$	00000	10111	01101	11010
000	00000	00000	10111	01101	11010
111	10000	10000	00111	101101	01010
101	01000	01000	11111	00101	10010
100	00100	00100	10011	01001	11110
010	00010	0010	10101	01111	11000
001	00001	00001	10110	01100	11011
011	00011	00011	10100	01110	11001
110	00110	00110	10001	01011	11100

【例 5.8】某(5，2)系统线性码的生成矩阵是

$$\boldsymbol{G}=\begin{bmatrix}1 & 0 & 1 & 1 & 1\\ 0 & 1 & 1 & 0 & 1\end{bmatrix}$$

接收码是 $\boldsymbol{R}=(10101)$。求发码的估值 $\hat{\boldsymbol{C}}$。

【解】

① 以信息组 $\boldsymbol{m}=[0\quad 0]$、$[0\quad 1]$、$[1\quad 0]$、$[1\quad 1]$及已知的 $\boldsymbol{G}$，代入 $\boldsymbol{C}=\boldsymbol{mG}$，求出 4 个许用码字

$$\boldsymbol{C}_1=[0\quad 0\quad 0\quad 0\quad 0]$$
$$\boldsymbol{C}_2=[0\quad 1\quad 1\quad 0\quad 1]$$
$$\boldsymbol{C}_3=[1\quad 0\quad 1\quad 1\quad 1]$$
$$\boldsymbol{C}_4=[1\quad 1\quad 0\quad 1\quad 0]$$

② 由 $\boldsymbol{G}\rightarrow\boldsymbol{H}$，即

$$\begin{aligned}\boldsymbol{H}&=[\boldsymbol{P}^{\mathrm{T}}\quad \boldsymbol{I}]\\&=\begin{bmatrix}1 & 1 & 1 & 0 & 0\\ 1 & 0 & 0 & 1 & 0\\ 1 & 1 & 0 & 0 & 1\end{bmatrix}\\&=[\boldsymbol{H}_1\quad \boldsymbol{H}_2\quad \boldsymbol{H}_3\quad \boldsymbol{H}_4\quad \boldsymbol{H}_5]\end{aligned}$$

③ 求标准阵列。因为 $\boldsymbol{S}^{\mathrm{T}}=\boldsymbol{HE}^{\mathrm{T}}=\sum_{i=1}^{n}e_i\boldsymbol{H}_i$，所以无差错时，$\boldsymbol{S}=\boldsymbol{0}$；当只有一个码元发生差错时，$\boldsymbol{S}^{\mathrm{T}}$ 等于 $\boldsymbol{H}$ 列对应的；当两个以上码元发生差错时，$\boldsymbol{S}^{\mathrm{T}}$ 等于对应列的矢量和。

由于 $n-k=3$ 所以伴随式共有 $2^{n-k}=8$ 种，而差错图案中表示无差错的一种，表示一个差错的图样有 $C_5^1=5$ 种。8 个伴随式中对应这 6 个重量最小的差错图案后还多出 2 个伴随式，让它们对应两个差错为 2 的图样。差错为 2 的图样共有 $C_5^2=10$ 种，从中选出两个。选的方法可有多种，不是唯一的。

将 $\boldsymbol{E}=(00000),(10000),(01000),(00100),(00010),(00001)$ 代入 $\boldsymbol{S}^{\mathrm{T}}=\sum_{i=1}^{n}e_i\boldsymbol{H}_i$，求出对应的 $\boldsymbol{S}$ 为(000)，(111)，(101)，(100)，(010)，(001)。剩下的伴随式中，(011)对应的图样有 2^k 个（$\boldsymbol{S}^{\mathrm{T}}=\sum_{i=1}^{n}e_i\boldsymbol{H}_i$ 是 $n-k$ 个方程，n 个未知数，因此每个未知数有 2^k 个解），即(00011)，(10100)，(01110)，(11001)，其中(00011)和(10100)并列重量最轻，任选其中一个比如(00011)。同样，伴随式(110)选一个对应的最轻的差错图案(00110)。

根据以上讨论，画出标准阵列如表 5.5 所示。

④ 由 $\boldsymbol{R}\rightarrow\hat{\boldsymbol{C}}$。例如，若 $\boldsymbol{S}=\boldsymbol{R}\boldsymbol{H}^{\mathrm{T}}=[0\quad 1\quad 0]$，查出对应的差错图案为 $\boldsymbol{E}=[0\quad 0\quad 0\quad 1\quad 0]$。所以，$\hat{\boldsymbol{C}}=\boldsymbol{R}+\boldsymbol{E}=[1\quad 0\quad 1\quad 0\quad 1]+[0\quad 0\quad 0\quad 1\quad 0]=[1\quad 0\quad 1\quad 1\quad 1]$。

进一步分析可知，该码的 $d_{\min}=3$，纠错能力 $u=1$，所以译码阵列中只有前 6 行具有唯一性、可靠性。而第 7 和第 8 行是表示有 2 个差错的图案，已超出了 $u=1$ 的纠错能力，译码已不可靠。第 7、8 行在选择差错图案时有多种选法，选法不同，造成最终的译码结果就会不同。

由于表示单个错误的错误图样的伴随式就等于 $\boldsymbol{H}$ 矩阵中错误码元的对应列，为了使不同错误码元对应不同的伴随式以便译码，就应使 $\boldsymbol{H}$ 中的 n 列互不相同且不能为 0（若某列为 0 就不能表示对应码元出现错误的情况）。

伴随式的个数 2^{n-k} 与 n、k 及纠错能力 u 之间满足以下关系。

【定理 5-5】若 (n,k) 线性分组码能够纠正 u 个错误，则其校验位的数目必须满足

$$2^{n-k}\geqslant\sum_{i=0}^{u}C_n^i \tag{5.10}$$

【证明】

由于产生 $i(i\leqslant u)$ 个错误的错误图样有 C_n^i 种，能够产生不多于 u 个错误的错误图样，共 $C_n^0+C_n^1+\cdots+C_n^u=\sum_{i=0}^{u}C_n^i$ 个。

而 $n-k$ 位校验元有 2^{n-k} 种不同的组合，即有 2^{n-k} 个伴随式。如果某线性分组码能够纠正 u 个错误，则 $\sum_{i=0}^{u}C_n^i$ 个错误图样都应该有一个对应的伴随式，即伴随式的数目须满足条件

$$2^{n-k}\geqslant\sum_{i=0}^{u}C_n^i$$

证毕。

式(5.10)等号成立时的线性分组码称为完备码，即完备码的伴随式数目不多不少恰好和不大于 u 个差错的图案数目相等，这时校验位得到最充分的利用，所有的伴随式都唯一可靠地对应一个差错图案。

从多维矢量空间的角度来看完备码。假定我们围绕每个码字 $\boldsymbol{C}_i$，有一个半径为 u 的球，每个球内包含了与该码字的汉明距离小于等于 u 的所有接收码字 $\boldsymbol{R}$ 的集合，所有落在这个球内的接收码字都被译为该发送码字。这样在这个半径为 $u=[(d_{\min}-1)/2]$ 的球内的接收码字数为 $\sum_{i=0}^{u}C_n^i$。因为有 2^k 个可能发送的码字，也就有 2^k 个不相重叠的半径为 u 的球。因

为包含在这 2^k 个球中的码字总数不会超过 2^n 个可能的接收码字，所以一个能纠 u 个差错的码必然满足不等式

$$2^k \times \sum_{i=0}^{u} C_n^i \leqslant 2^n$$

即

$$2^{n-k} \geqslant \sum_{i=0}^{u} C_n^i$$

当 $u=1$ 时，有 $2^{n-k} \geqslant C_n^0 + C_n^1 = 1+n$。

当式(5.10)等号成立时，表示所有的接收码字都落在 2^k 个球内而球外没有一个码字。这就是完备码。完备码具有下述特性：

① 每个接收码字都落在这些球中之一，因此接收码字与发送码字的距离至多为 u。

② 所有差错数小于等于 u 的接收码字都能得到纠正，而差错数大于等于 $u+1$ 的接收码字因为落在另一个球内被纠为其他发送码字。

完备码并不多见，我们知道的有 $u=1$ 的汉明码、$u=3$ 的高莱码以及$(n, 1)$（n 为奇数）简单重复码等。

5.3.4 几种重要的线性分组码

下面重点介绍线性分组码中理论和实用价值都比较大的汉明码和循环码。

1. 汉明码

汉明码是是香农的信道编码定理提出后最早发现的码，有二进制的，也有非二进制的。我们讨论二进制的汉明码。

汉明码是能够纠正一个错误的完备码，因此它的码长 n 和信息位数 k 服从以下规律：

$$(n,k) = (2^m - 1, 2^m - m - 1)$$

其中，$m = n - k$ 为校验位数。

当 $m=3,4,5,\cdots$时，有$(7, 4)$，$(15, 11)$，$(31, 26)$，$(63, 57)$，$(127, 120)$，$(255, 247)\cdots$汉明码。一个(n, k)汉明码的校验矩阵有 $n-k$ 行和 n 列。二进制时，$n-k$ 个码元所能组成的列矢量总数（全零矢量除外）是 $2^{n-k}-1$，恰好与校验矩阵的列数 $n=2^m-1$ 相等，因此只要把 $n-k$ 码元组成的列矢量按二进制数大小顺序从左到右排列，就可以得到它的 $\boldsymbol{H}$。这样得到的 $\boldsymbol{H}$ 是非系统码。当发生单个错误时，伴随式是 $\boldsymbol{H}$ 中与错误位置对应的列，并且伴随式的二进制数的值就是错误位置的序号，因此它的编译码规则都很简单。

如果想得到系统形式的 $\boldsymbol{H}$，则可以通过列置换把非系统形式的 $\boldsymbol{H}$ 变成系统形式的 $\boldsymbol{H}$。

【例 5.9】构造一个 $m=3$ 的二元$(7, 4)$汉明码。

【解】

所谓构造，就是求一个$(7, 4)$汉明码的生成矩阵。先利用汉明码的特性构造一个校验矩阵 $\boldsymbol{H}$，再通过列置换将它变为系统形式。

$$\boldsymbol{H} = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix} \xrightarrow{\text{列置换}} \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} = [\boldsymbol{P}^{\mathrm{T}} \quad \boldsymbol{I}_3]$$

$$\longrightarrow \boldsymbol{G}=(\boldsymbol{I}_4 \quad \boldsymbol{P})=\begin{bmatrix}1&0&0&0&0&1&1\\0&1&0&0&1&0&1\\0&0&1&0&1&1&0\\0&0&0&1&1&1&1\end{bmatrix}$$

由于 $\boldsymbol{G}$ 中包含了单位阵 $\boldsymbol{I}_4$，矩阵的秩是 4，所以矩阵 $\boldsymbol{G}$ 的 4 行是 4 个线性无关的基底，可以张成一个包含 $2^4=16$ 个码字的码空间。

必须指出，完备码是标准阵列最规则因而译码最简单的码，但并不一定是纠错能力最强的码，因为它不一定是极大最小距离码，即不满足 $d_{\min}=n-k+1$。$(7,4)$汉明码的 $d_{\min}=3$，而同样 n 和 k 的极大最小距离码 $d_{\min}=4$。

如果给(n,k)汉明码添加一位奇偶校验位，可得到一个 $d_{\min}=4$ 的$(n+1,k)$扩展汉明码，码长为 $n+1$，校验位为 $n-k+1$。这时

$$\boldsymbol{H}'=\begin{bmatrix}&&&0\\&&&0\\&\boldsymbol{H}&&\vdots\\&&&0\\1&1&\cdots&1\end{bmatrix}$$

其中，$\boldsymbol{H}$ 为$(n-k)\times n$ 矩阵，$\boldsymbol{H}'$为$(n-k+1)\times(n+1)$矩阵。

这个码能纠正单个错误外还能发现两个错误。因为当码字在传输过程中发生两个错误时，其伴随式为对应校验矩阵中的两列之和。为了能够发现两个错误，必须使得校验矩阵的任意两列之和不为其他列，即要求校验矩阵中的任意三列线性无关，即码字的最小距离为 4。

$$d_{\min}=4=1+2+1$$

即 $u=1$，$l=2$。当出现一位错误时，伴随式为$\boldsymbol{H}'$的某一列，最后一位数为 1。当出现两个错误时，伴随式为某两列之和，最后一位为 0。因为它与$\boldsymbol{H}'$中的任何一列都不相同，所以可与单个错误的伴随式区别开来，故可以检查两个错误。

在同样的纠错能力下，汉明码的码率是最高的，即

$$R=\frac{2^m-1-m}{2^m-1}=1-\frac{m}{2^m-1}$$

当 m 很大时，$R\to 1$。

2. 循环码

循环码是线性分组码的一个子类，它具有完整的代数结构，编译和译码可以用具有反馈联级的移位寄存器来实现。它满足循环移位特性：码 $\boldsymbol{C}$ 中任何一个码字的循环移位仍是码 $\boldsymbol{C}$ 中的一个码字。

【定义 5-6】对于一个(n,k)线性分组码，若某一码字为 $\boldsymbol{C}=(C_{n-1},C_{n-2},\cdots,C_2,C_1,C_0)$，该码字向左循环一位后为 $\boldsymbol{C}^{(1)}=(C_{n-2},C_{n-3},\cdots,C_1,C_0,C_{n-1})$，向左循环移动 $n-1$ 位后为 $\boldsymbol{C}^{(n-1)}=(C_0,C_{n-1},C_{n-2},\cdots,C_1)$。若 $\boldsymbol{C}^{(i)}(i=1,2,\cdots,n-1)$均为码字，则称这个$(n,k)$线性分组码为循环码。这里循环移位也可以定义为向右移位。

一般(n,k)线性分组码的 k 个基底之间不存在规则的联系，因此我们需用 k 个基底组成

生成矩阵来表示一个码的特征。而循环码的 k 个基底可以是同一个基底循环 k 次得到，因此用一个基底就可以表示一个码的特征。我们可以用不大于 $n-1$ 次的码多项式 $C(x)$ 来表示一个码字：

$$C(x)=c_{n-1}x^{n-1}+\cdots+c_2x^2+c_1x+c_0$$

这里，码元序号从 $0\rightarrow n-1$ 而不用 $1\rightarrow n$ 是为了在以后的多项式运算中系数序号与 x 的幂次一致。

循环码的循环特性可以用码多项式表示为；

移 1 位：$C^{(1)}(x)=xC(x)=C_{n-2}x^{n-1}+\cdots+C_1x^2+C_0x+C_{n-1}$

移 2 位：$C^{(2)}(x)=x^2C(x)=C_{n-3}x^{n-1}+\cdots+C_0x^2+C_{n-1}x+C_{n-2}$

$\vdots$

移 $n-1$ 位：$C^{(n-1)}(x)=x^{n-1}C(x)=C_0x^{n-1}+\cdots+C_3x^2+C_2x+C_1$

$C(x)$ 移 n 位后又回到 $C(x)$，一个码字的移位最多能得到 n 个码字，因此循环码码字的循环仍是码字并不意味着循环码可以仅从一个码字循环而得。一个 (n,k) 循环码有 2^k 个码字，它们都是同一基底的线性组合。根据线性码空间的封闭性，码字的线性组合仍是码字。

在 2^k 个码字的码多项式中取一个次数最低即 $n-k$ 次的多项式作为生成多项式，用 $g(x)$ 表示。可以证明，$g(x)$ 是码多项式中唯一一个 $n-k$ 次的多项式且常数项不为 0，即 g_0 及 g_{n-k} 均为 1。由生成多项式 $g(x)$ 可以得到循环码的生成矩阵。因为 $x^ig(x)\ (i=0,1,2,\cdots,k-1)$ 均是码字且线性无关，故可用来构成一个生成矩阵。

令 $g(x)=x^{n-k}+\cdots+g_2x^2+g_1x+1$，有

$$\boldsymbol{G}(x)=\begin{bmatrix}x^{k-1}g(x)\\ xg(x)\\ g(x)\end{bmatrix}$$

$$=\begin{bmatrix}1 & g_{n-k-1} & g_{n-k-2} & \cdots & g_1 & 1 & 0 & 0 & \cdots & 0\\ 0 & 1 & g_{n-k-1} & g_{n-k-2} & \cdots & g_1 & 1 & 0 & \cdots & 0\\ 0 & 0 & 1 & g_{n-k-1} & g_{n-k-2} & \cdots & g_1 & 1 & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 0 & 1 & g_{n-k-1} & g_{n-k-2} & \cdots & \cdots & 1\end{bmatrix}$$

$g(x)$ 的系数是降幂排列。

由生成多项式 $g(x)$ 和信息多项式可以得到循环码的码多项式：

$$\begin{aligned}C(x)&=(m_{k-1}m_{k-2}\cdots m_1m_0)g(x)\\&=\sum_{i=0}^{k-1}m_ix^ig(x)\\&=m(x)g(x)\end{aligned}$$

每个码多项式 $C(x)$ 都是 $g(x)$ 的倍式，并且每个次数小于等于 $n-1$ 的 $g(x)$ 的倍式必是一个码多项式。

生成多项式 $g(x)$ 一定是 x^n+1 的因子，即 $x^n+1=g(x)h(x)$，这是用 $g(x)$ 构造循环码的充要条件。反过来，如果 $g(x)$ 是 x^n+1 的 $n-k$ 次因子，一定可以构造一个 (n,k) 循环码。这样可以保证码的循环移位特性，即码字的循环仍是码字。

构造(n,k)循环码的步骤如下：

① 对x^n+1进行因式分解，找出$n-k$次因式。

② 以该$n-k$次因式作为生成多项式，与不高于$k-1$次的信息多项式相乘得码多项式$C(x)=m(x)g(x)$。$C(x)$的次数不高于$(k-1)+(n-k)=n-1$次。

【例 5.10】构造一个$(7,4)$循环码。

【解】

(1) 对(x^7+1)进行因式分解得：$x^7+1=(x+1)(x^3+x^2+1)(x^3+x+1)$。3次因式有两个$(x^3+x^2+1)$和$(x^3+x+1)$，均可以作为$(7,4)$循环码的生成多项式，选择不同的$g(x)$会得到不同的$(7,4)$循环码。

(2) 选$g(x)=x^3+x^2+1$，信息多项式共有$2^k=16$种可能的组合，对应16个码字。利用$C(x)=m(x)g(x)$可得到16个码字。

例如，$\boldsymbol{m}=[0\quad 1\quad 1\quad 0]$对应码字：

$$\begin{aligned}C(x)&=m(x)g(x)\\&=(m_3x^3+m_2x^2+m_1x+m_0)g(x)\\&=x^5+x^3+x^2+x\rightarrow(0101110)\end{aligned}$$

表5.6是该$(7,4)$循环码的码字。可以看出，任何码字的循环仍然是码字，整个码组有4组码字的循环，但都是$g(x)=x^3+x^2+1$的线性组合。

表 5.6 (7,4)循环码

信息比特 $m_3m_2m_1m_0$	码字（循环1） $C_6C_5C_4C_3C_2C_1C_0$	信息比特 $m_3m_2m_1m_0$	码字（循环2） $C_6C_5C_4C_3C_2C_1C_0$	信息比特 $m_3m_2m_1m_0$	码字（循环3和4） $C_6C_5C_4C_3C_2C_1C_0$
0001	0001101	0011	0010111	0000	0000000
0010	0011010	0110	0101110	1111	1111111
0100	0110100	1100	1011100		
1000	1101000	0101	0111001		
1101	1010001	1010	1110010		
0111	0100011	1001	1100101		
1110	1000110	1111	1001011		

由本例可以验证循环码码字的循环仍是码字，码字的线性组合也仍是码字。

$x^n+1=g(x)h(x)$中的$h(x)$称为该循环码的一致校验多项式，其阶次为k。$h(x)$的校验作用表现为：任何码多项式$C(x)$与$h(x)$的乘积一定等于0（模x^n+1），而非码字与$h(x)$的乘积必不为0，因为

$$C(x)h(x)=m(x)g(x)h(x)=m(x)(x^n+1)=0\quad \mod\quad (x^n+1)$$

在$x^n+1=g(x)h(x)$的因式分解中，$g(x)$和$h(x)$处于同等地位。既然可以用$g(x)$生成一个循环码，也就可以用$h(x)$生成另一个循环码。此时$h(x)$用于生成多项式，而$g(x)$用于一致校验多项式。由$g(x)$生成的(n,k)循环码和$h(x)$生成的$(n,n-k)$循环码互为对偶码。

由校验多项式可以得到校验矩阵。

令$h(x)=h_kx^k+\cdots+h_1x+h_0=\sum_{i=0}^{k}h_ix^i$，校验矩阵为

$$\boldsymbol{H}=\begin{bmatrix} h_0 & h_1 & \cdots & h_k & 0 & 0 & \cdots & 0 \\ 0 & h_0 & h_1 & \cdots & h_k & 0 & \cdots & 0 \\ 0 & 0 & h_0 & h_1 & \cdots & h_k & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \cdots & h_0 & h_1 & \cdots & h_k \end{bmatrix}$$

记 $h^*(x)=h_0x^k+\cdots+h_{k-1}x+h_k$ 为 $h(x)$ 的倒多项式，即把 $h(x)$ 的系数倒过来排列，则校验矩阵可以表示成：

$$\boldsymbol{H}(x)=\begin{bmatrix} x^{n-k-1}h^*(x) \\ \vdots \\ xh^*(x) \\ h^*(x) \end{bmatrix}$$

可以证明，$x^ih^*(x)(i=1,2,\cdots,n-k-1)$ 是线性无关的向量，并且 $\boldsymbol{GH}^{\mathrm{T}}=\boldsymbol{0}$。

这样得到循环码的生成矩阵和校验矩阵均不是系统形式的。通过矩阵的初等变换，可以将其变成系统形式的 $\boldsymbol{G}$ 和 $\boldsymbol{H}$。

由循环码的生成矩阵 $\boldsymbol{G}$ 和校验矩阵 $\boldsymbol{H}$ 的表示式可知，$\boldsymbol{G}$ 和 $\boldsymbol{H}$ 有一定的关系，由 $\boldsymbol{G}$ 可以得到 $\boldsymbol{H}$，反之由 $\boldsymbol{H}$ 可以得到 $\boldsymbol{G}$。

令 $\boldsymbol{C}$ 为发送码字序列，$\boldsymbol{E}$ 为错误图样序列，则接收序列为 $\boldsymbol{R}=\boldsymbol{C}+\boldsymbol{E}$。定义接收序列的伴随式为 $\boldsymbol{S}=\boldsymbol{RH}^{\mathrm{T}}$，由于 $\boldsymbol{CH}^{\mathrm{T}}=\boldsymbol{0}$，则有 $\boldsymbol{S}=\boldsymbol{RH}^{\mathrm{T}}=[\boldsymbol{C}+\boldsymbol{E}]\boldsymbol{H}^{\mathrm{T}}=\boldsymbol{EH}^{\mathrm{T}}$，因此接收序列的伴随式包含了错误图样信息，可用于译码。

【例 5.11】已知(7, 4)循环码的生成多项式 $g(x)=x^3+x+1$，求它的系统形式的生成矩阵。

【解】

$$\boldsymbol{G}=\begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \xrightarrow[\text{把第三、四行加到第一行}]{\text{将第四行加到第二行}} \boldsymbol{G}=\begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

循环码的译码是利用伴随多项式来完成的。

设发送码的多项式 $C(x)=\sum_{i=0}^{n-1}c_ix^i$，错误图样多项式 $e(x)=\sum_{i=0}^{n-1}e_ix^i$，接收码多项式 $R(x)=\sum_{i=0}^{n-1}r_ix^i$，则有 $R(x)=C(x)+e(x)$。

设 $g(x)$ 为码的生成多项式，因为发送码多项式 $C(x)$ 能被 $g(x)$ 除尽，所以

$$\frac{R(x)}{g(x)}=\frac{C(x)+e(x)}{g(x)}=\frac{e(x)}{g(x)}$$

定义伴随多项式（简称伴随式）为

$$s(x)=\frac{e(x)}{g(x)}=e(x)\mod g(x)$$

若无错误传输，则 $s(x)=0$，否则 $s(x)\neq0$。

由伴随多项式，可得到发送码多项式的估值：

$$\hat{C}(x)=R(x)+e(x) \mod g(x)$$

因为 $g(x)$ 的次数为 $n-k$，$e(x)$ 的次数为 $n-1$，则伴随式 $s(x)$ 的最高次数为 $n-k-1$，即 $s(x)$ 共有 $n-k$ 项，故有 2^{n-k} 种可能的表示式，即有 2^{n-k} 个伴随式。若 $2^{n-k}\geqslant n+1$，则具有至少纠正一位错误的能力。

【例 5.12】已知(7,4)循环码的生成多项式 $g(x)=1+x+x^3$，若接收码的最高位码元发生错误，求其伴随多项式；若接收码字为 0001110，求发送码字。

【解】

求 $s(x)$ 的计算实际上是对 $g(x)$ 做除法求余运算。

已知 $\boldsymbol{E}=[1\ 0\ 0\ 0\ 0\ 0\ 0]$，对应的错误图样多项式为 $e(x)=x^6$，则

$$s(x)=\frac{e(x)}{g(x)}=\frac{x^6}{1+x+x^3}=(1+x^2) \mod g(x)$$

对应的伴随式为 $\boldsymbol{S}=[1\ 0\ 1]$。

接收码字(0001110)的码多项式为

$$R(x)=x^3+x^2+x$$

$$s(x)=\frac{e(x)}{g(x)}=\frac{R(x)}{g(x)}=\frac{x^3+x^2+x}{x^3+x+1}=(x^2+1) \mod g(x)$$

根据 $s(x)$ 从译码表中找出对应的错误图样多项式为 x^6，可得到发送码字的估值为

$$\begin{aligned}\hat{C}(x)&=R(x)+e(x)\\&=x^3+x^2+x+x^6=(1001110)\end{aligned}$$

循环码是线性分组码中非常重要的一个子类，要设计一个码率 $R=\frac{k}{n}$ 的循环码只要将 x^n+1 解出一个 $n-k$ 次因式 $g(x)$，就可以生成一个 (n,k) 循环码。目前，有实用价值的纠错码大部分都属于循环码的范围。比如，在无线信道上应用最广泛的 BCH、RS 码等。

BCH 码是循环码的一个重要子类，具有纠多个错误的能力，BCH 码有严密的代数理论，是目前研究最透彻的一类码。BCH 码的生成多项式与最小码距之间有密切的关系，人们可以根据所要求的纠错能力很容易构造出 BCH 码，它们的译码器也容易实现，是线性分组码中应用最普遍的一类码。

RS 码是多进制 BCH 码，具有很强的纠错能力，随机错误、突发错误都能纠正，特别是短的和中等码长的性能接近理论值，同时构造方便，编码简单，因此是目前最有效、应用最广泛的差错控制编码，在 CD、DVD、蓝光光盘、宽带无线接入技术 WiMAX 中都得到应用。

5.3.5 卷积码*

前面研究过的各种分组码都是将序列切割后分组进行编译码。信息序列被分组后分组之间的相关信息就损失了，并且分组长度越小，损失的信息就越多。如果把分组长度取得很大，则译码复杂度随之呈指数上升。于是我们考虑在码长 n 有限的情况下，将若干个分组的相关性消息添加到码字里，从而等效地增加码长。译码时，利用前后码字的相关性，将前面的译码信息反馈到后面，作为译码参考。这样编码器在某个时间段产生的 n 个码元，不但取

决于该时间段进入编码器的 k 个信息位，而且与前面的 $N-1$ 个时间段内的信息组有关。这就是树码或称链码。

卷积码是树码中最重要的一类，它的码字与 N 个时间段的信息组的映射关系是时不变的线性关系，卷积码与分组码相似，具有纠正随机错误、突发错误或同时纠正这两类错误的能力。通常它更适用于前向纠错，因为其纠错性能对于许多实际情况常优于分组码，而且设备较简单，可用移位寄存器来完成编解码。

卷积码也可以用生成矩阵和校验矩阵来研究它的编解码。

下面以(3，1)卷积码为例，讨论卷积码的生成矩阵和校验矩阵。

把给定的信息序列($m_1 \quad m_2 \quad m_3 \quad \cdots$)进行分组，使每组只包含一个信息位 m_1，校验位有两位 p_{i_1} 和 p_{i_2}，对应的码序列为($m_1p_{11}p_{12} \quad m_2p_{21}p_{22} \quad m_3p_{31}p_{32} \quad \cdots$)。假设校验位与信息位满足以下关系：

$$p_{i1} = m_i + m_{i-1} + m_{i-3}$$

$$p_{i2} = m_i + m_{i-1} + m_{i-2}$$

当前的校验位与当前的信息位和过去的三个信息位有关，且满足线性关系。某信息位影响 4 个分组，即该卷积码的约束长度为 4。

考察编码器的输入输出之间的关系。令

$$\boldsymbol{m} = [m_1 \quad m_2 \quad m_3 \quad m_4 \quad \cdots]$$

$$\boldsymbol{C} = [m_1p_{11}p_{12} \quad m_2p_{21}p_{22} \quad m_3p_{31}p_{32} \quad m_4p_{41}p_{42} \quad \cdots]$$

把监督位与信息位的关系代入，得

$$\boldsymbol{C} = [m_1m_1m_1m_2(m_1+m_2)(m_1+m_2)m_3(m_2+m_3)(m_1+m_2+m_3)$$
$$m_4(m_1+m_3+m_4)(m_2+m_3+m_4)\cdots]$$

已知 $\boldsymbol{m}$ 可以利用下式来生成 $\boldsymbol{C}$：

$$\boldsymbol{C} = [m_1 \quad m_2 \quad m_3 \quad m_4 \cdots] \begin{bmatrix} 111 & 011 & 001 & 010 & 000 & \cdots \\ 000 & 111 & 011 & 001 & 010 & \cdots \\ 000 & 000 & 111 & 011 & 001 & \cdots \\ 000 & 000 & 000 & 111 & 011 & \cdots \\ 000 & 000 & 000 & 000 & 111 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

$$= \boldsymbol{mG}$$

上式中的矩阵 $\boldsymbol{G}$ 被称为卷积码的生成矩阵，它是一个有头无尾的半无穷矩阵。

生成矩阵还可以写成分块矩阵的形式：

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{P}_1 & \boldsymbol{0} & \boldsymbol{P}_2 & \boldsymbol{0} & \boldsymbol{P}_3 & \boldsymbol{0} & \boldsymbol{P}_4 & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ & & \boldsymbol{I} & \boldsymbol{P}_1 & \boldsymbol{0} & \boldsymbol{P}_2 & \boldsymbol{0} & \boldsymbol{P}_3 & \boldsymbol{0} & \boldsymbol{P}_4 & \cdots \\ & & & & \boldsymbol{I} & \boldsymbol{P}_1 & \boldsymbol{0} & \boldsymbol{P}_2 & \boldsymbol{0} & \boldsymbol{P}_3 & \cdots \\ & & & & & & \mathrm{I} & \boldsymbol{P}_1 & \boldsymbol{0} & \boldsymbol{P}_2 & \cdots \\ & & & & & & & & \boldsymbol{I} & \boldsymbol{P}_1 & \cdots \\ & & & & & & & & & & \vdots \end{bmatrix}$$

$\boldsymbol{I}$ 为 $k \times k = 1 \times 1$ 阶单位阵，$\boldsymbol{0}$ 为 $k \times k = 1 \times 1$ 阶全 0 方阵，$\boldsymbol{P}_i$ 为 $k \times (n-k) = 1 \times 2$ 阶矩阵。

$$\boldsymbol{P}_1=[1\quad 1]\qquad \boldsymbol{P}_2=[1\quad 1]\qquad \boldsymbol{P}_3=[0\quad 1]\qquad \boldsymbol{P}_4=[1\quad 0]$$

可以看到，生成矩阵每一行都相同，只不过每行是上一行向右移动 3 列。

输入信息序列 $\boldsymbol{m}=[0\ 1\ 0\ 0\ 1\ 0\ 1\ 0\ 1\ 1\ \cdots]$，对应的码字为 $\boldsymbol{C}=\boldsymbol{mG}=[0\ 0\ 0\ 1\ 1\ 1\ 0\ 1\ 1\ 0\ 0\ 1\ \cdots]$。对于所生成的码序列，每 3 个数字组成一个码字，一个码字包括一个信息位和两个校验位。

由于生成矩阵每一行都相同，所以矩原矩阵完全可以由第一行确定。把第一行 $\boldsymbol{G}_0=[\boldsymbol{I}\quad \boldsymbol{P}_1\quad \boldsymbol{0}\quad \boldsymbol{P}_2\quad \boldsymbol{0}\quad \boldsymbol{P}_3\quad \boldsymbol{0}\quad \boldsymbol{P}_4\quad \cdots]$称为基本生成矩阵。这里 $\boldsymbol{G}_0$ 的设定具有一般性。以时刻 i 为基准（一般可将 i 理解为编码时刻或当前时刻），这个时刻在编码过程中是不断向前移动的。设编码器的初始状态为零（移位寄存器清 0），则随着一个个 k 比特信息组的输入，编码器不断地输出码字。

由校验位与信息位的关系还可以确定基本一致校验矩阵。

把有约束关系的 4 个码字写成

$$\boldsymbol{C}_0=[m_{i-3}p_{i-3,1}p_{i-3,2}\quad m_{i-2}p_{i-2,1}p_{i-2,2}\quad m_{i-1}p_{i-1,1}p_{i-1,2}\quad m_ip_{i_1}p_{i_2}]$$

由 $\begin{cases}m_{i-3}+m_{i-1}+m_i+p_{i_1}=0\\ m_{i-2}+m_{i-1}+m_i+p_{i_2}=0\end{cases}$，可得

$$\begin{bmatrix}100 & 000 & 100 & 110\\ 000 & 100 & 100 & 101\end{bmatrix}\boldsymbol{C}_0^{\mathrm{T}}=\boldsymbol{0}$$

令

$$\boldsymbol{H}_0=\begin{bmatrix}100 & 000 & 100 & 110\\ 000 & 100 & 100 & 101\end{bmatrix}$$

$\boldsymbol{H}_0$ 称为该卷积码的基本一致校验矩阵。它可以判断有约束关系的 4 个接收码字是否是发送码字，与线性分组码的一致校验矩阵一样起着校验作用。由于输入序列 $\boldsymbol{m}$ 是一个半无限长序列，生成的卷积码也是一个半无限长的码序列。对这个半无限长的码序列 $\boldsymbol{C}=[m_1p_{11}p_{12}\ m_2p_{21}p_{22}\quad m_3p_{31}p_{32}\quad m_4p_{41}p_{42}\cdots]$的校验矩阵也是一个有头无尾的半无穷矩阵。

由校验位和信息位的关系可以得到

$$\begin{cases}m_1+p_{11}=0\\ m_1+p_{12}=0\end{cases}$$

$$\begin{cases}m_1+m_2+p_{21}=0\\ m_1+m_2+p_{22}=0\end{cases}$$

$$\begin{cases}m_2+m_3+p_{31}=0\\ m_1+m_2+m_3+p_{32}=0\end{cases}$$

$$\begin{cases}m_1+m_3+m_4+p_{41}=0\\ m_2+m_3+m_4+p_{42}=0\end{cases}$$

$$\begin{cases}m_2+m_4+m_5+p_{51}=0\\ m_3+m_4+m_5+p_{52}=0\end{cases}$$

$$\vdots$$

可以得到

$$\begin{bmatrix} 110 & 000 & \cdots & & & & \\ 101 & 000 & \cdots & & & & \\ 100 & 110 & 000 & \cdots & & & \\ 100 & 101 & 000 & \cdots & & & \\ 000 & 100 & 110 & 000 & \cdots & & \\ 000 & 100 & 101 & 000 & \cdots & & \\ 100 & 000 & 100 & 110 & 000 & \cdots & \\ 000 & 100 & 100 & 101 & 000 & \cdots & \\ 000 & 100 & 000 & 100 & 110 & 000 & \cdots \\ 000 & 000 & 100 & 100 & 101 & 000 & \cdots \\ & & & & & & \vdots \end{bmatrix} \cdot \boldsymbol{C}^{\mathrm{T}} = 0$$

记系数矩阵为 $\boldsymbol{H}$，称为该(3，1)卷积码的一致校验矩阵。

可以看到，这个有头无尾的半无穷矩阵每 3 列的结构相同，但后 3 列比前 3 列向下移 2 行。从第 7 行开始每 2 行结构相同。可以看出，这 2 行就是基本一致校验矩阵。

上式中的校验矩阵还可写为

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{P}_1^{\mathrm{T}} & \boldsymbol{I} & \cdots & & & & & & & & \\ \boldsymbol{P}_2^{\mathrm{T}} & \boldsymbol{0} & \boldsymbol{P}_1^{\mathrm{T}} & \boldsymbol{I} & \cdots & & & & & & \\ \boldsymbol{P}_3^{\mathrm{T}} & \boldsymbol{0} & \boldsymbol{P}_2^{\mathrm{T}} & \boldsymbol{0} & \boldsymbol{P}_1^{\mathrm{T}} & \boldsymbol{I} & \cdots & & & & \\ \boldsymbol{P}_4^{\mathrm{T}} & \boldsymbol{0} & \boldsymbol{P}_3^{\mathrm{T}} & \boldsymbol{0} & \boldsymbol{P}_2^{\mathrm{T}} & \boldsymbol{0} & \boldsymbol{P}_1^{\mathrm{T}} & \boldsymbol{I} & \cdots & & \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{P}_4^{\mathrm{T}} & \boldsymbol{0} & \boldsymbol{P}_3^{\mathrm{T}} & \boldsymbol{0} & \boldsymbol{P}_2^{\mathrm{T}} & \boldsymbol{0} & \boldsymbol{P}_1^{\mathrm{T}} & \boldsymbol{I} & \cdots \\ & & & & & & & & & & \vdots \end{bmatrix}$$

其中，$\boldsymbol{P}_i^{\mathrm{T}}$ 为 $(n-k)\times k = 2\times 1$ 维矩阵，$\boldsymbol{0}$ 为 $(n-k)\times(n-k) = 2\times 2$ 全 0 方阵，$\boldsymbol{I}$ 为 $(n-k)\times(n-k) = 2\times 2$ 维单位阵。

由以上生成矩阵和校验矩阵的讨论可以看到，它们与基本生成矩阵有密切的关系，以矩阵的方式描述了卷积码的卷积关系式。在卷积码的应用中，给定基本生成矩阵就可以确定卷积码的编码电路及其矩阵表达式。

卷积码的译码分为代数译码和概率译码。代数译码从码的代数结构出发，概率译码从信道的统计特性出发。门限译码是一种代数译码，算法简单、易于实现，比较实用。而序列译码和维特比最大似然译码则是概率译码。序列译码的延时是随机的，与信道干扰有关。维特比译码的运算时间是固定的，它对存储器级数较小的卷积码的译码很容易实现，目前被广泛地应用于现代信道中。

5.3.6 TCM 码、级联码、Turbo 码和 LDPC 码

网格编码调制（Trellis Coded Modulation，TCM）是将编码技术与调制技术结合起来，利用状态记忆和分集映射来增加码序列之间的距离。这种方法不需要增加信道带宽或者信号传输功率，而是利用信号集空间的冗余提高信息传输效率。

级联码是利用短码拼接成长码，使拼接后的码字具有短码的译码复杂度和长码的性能。

这是由于在一定的信道性能上，为了得到差错概率小的好码，往往需要增加码的长度，码长度的增加直接带来译码复杂度的提高。因此，级联码的思想给长码的性能带来很大的改善。

Turbo 码实际上是一种并行级联卷积码（Parallel Concatenated Convolutional Code），是由 C. Berrou、A. Glavieux 和 P. Thitimajshiwa 于 1993 提出的。Turbo 码由两个递归系统卷积码（Rescursive System Code，RSC），通过交织器，以并行级联的方式结合而成。这种方案采用反馈迭代译码方式，真正发掘了级联码的潜力，并以其类似于随机的编译码方式，突破了最小距离的短码设计思想，使它更加逼近了理想的随机码的性能。仿真结果表明，该编码方式有着极强的纠错能力，是目前所知的最为高效的编码方式之一。如果采用大小为 65535 的随机交织器，并且进行 18 次选代，则在信噪比≥0.7dB 时，码率为 0.5 的 Turbo 码在加性高斯自噪声（AWGN）信道上的误比特率 BER≤10^{-8}e，达到了近香农限的性能。无论是在高斯白噪声信道还是衰落信道中，Turbo 码都能取得良好的误码率性能。由于其性能远远超过了其他编码方式，因此得到了广泛的关注和发展，并对编码理论和研究方法产生了深远的影响。

低密度奇偶校验码（Low Density Parity Check Code，LDPC）是麻省理工学院 Robert Gallager 于 1962 年在博士论文中提出的一种具有稀疏生成矩阵和校验矩阵的线性分组码。它的性能逼近香农限，且描述和实现简单，易于进行理论分析和研究，译码简单且可实行并行操作，适合硬件实现。LDPC 几乎适用于所有信道，因此成为编码界近年来的研究热点。美国航天局已经将 Turbo 码和 LDPC 码列为深空通信的技术规范。

动手实践 5.1：Hamming(7,4)编译码器

Hamming(7,4)码的生成矩阵为

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

其校验矩阵为

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

（1）输入：长度为 4 的任意二进制序列。

（2）输出：输入数据经 Hamming(7,4)编码器编码之后，通过一个 BSC 信道（错误概率为 0.1）传输后，再经过 Hamming(7,4)译码器译码输出，得到信宿端的长度为 4 的二进制序列。

动手实践 5.2：通信系统仿真

这是一个综合性的大型实验，通过搭建一个包括信源、信源编译码器、信道、信道编译码器等各模块在内的仿真通信系统，使读者能够加深对本课程各个重点章节的理解，更好地掌握通信的本质意义。

说明：由于搭建一个完整通信系统的工作量较大，所以本实验可以使用 Matlab 等仿真工具。下面分别描述系统中各模块的需求。

① 离散信源：能以指定的概率分布$(p,1-p)$产生 0、1 符号构成的二进制信源符号序列。

② 信源编码器：信源编码器的输入是上一步产生的二进制符号序列。要求：能选择使用无编码（直通）、二进制香农编码、二进制霍夫曼编码、二进制费诺编码这四种信源编码方式中的任何一种。

在上一步指定信源的概率分布后，就可以马上生成这三种编码的码表，实际的编码工作只是查表而已。当然，直接对上一步指定的信源进行编码是不合适的，需要先进行信源的扩展，换句话说，需要确定信源分组的长度。这个长度 N 也是本系统的一个重要参数，是在系统运行之前由用户输入的。

③ 信道编码器：信道编码器的输入是信源编码器输出的二进制符号序列。编码方式要求能选择使用无编码、3 次重复编码、Hamming(7, 4)码这三种信道编码方式中的任何一种。

信道编码器是个简单的一一对应的函数转换模块，没有额外的控制参数，可以事先实现这三种编码器，统一其输入输出格式，运行时按照指定的类型直接使用即可。

④ 信道：其输入是信道编码器输出的二进制符号序列。经过传输后输出被噪声干扰和损坏了的二进制符号序列。要求能够模拟理想信道、给定错误概率 p 的 BSC 以及给定符号 0、1 各自错误概率 p、q 的任意二进制信道。

⑤ 信道译码器：由于信源经过信源编码器和信道编码器后的统计特性难以明确给出，所以此时理想译码器准则无法实施。因此，根据第④步给出的信道统计特性，采用极大似然译码准则进行译码。

⑥ 信源译码器：在第②步确定信源编码器后，即可同时确定信源译码器。信源译码器的工作仅仅是简单的查表。

要求：

（1）输入：各模块的相关参数。

（2）输出：信源产生的原始符号序列、信源译码器输出的符号序列、信道编码后的信息传输效率、整个通信过程的误比特率（BER）以及信道编译码过程中产生的误码率(BLER)。

提示：

（1）本实验中的信源模块部分都会用到随机数的产生。各种编程语言基本都提供了这个功能。

（2）Matlab 是一个优秀的系统仿真软件，而 Simulink 是 Matlab 中最著名的通信工具箱。上述实验要求中的很多功能由 Matlab 或 Simulink 已经实现并提供了方便的调用接口。例如二进制对称信道，在 Matlab 中就有一个 bsc() 函数完成了这个功能。同学们在设计、开发这个实验前应该花一些时间先熟悉 Matlab 及 Simulink。

习　题　5

5.1　一个快餐店只提供汉堡包和牛排，当顾客进店以后只须向厨房喊一声“B”或“Z”

就表示他点的是汉堡包或牛排。不过通常 8% 的概率厨师可能会听错。一般来说进店的顾客 90% 会点汉堡包，10% 会点牛排。问：

（1）这个信道的信道容量

（2）每次顾客点菜时提供多少信息？

（3）在这个信道可不可能正确地传递顾客点菜的信息？

5.2 设有一离散无记忆信道，其信道矩阵为

$$\boldsymbol{P}=\begin{bmatrix}\frac{1}{2} & \frac{1}{3} & \frac{1}{6}\\ \frac{1}{6} & \frac{1}{2} & \frac{1}{3}\\ \frac{1}{3} & \frac{1}{6} & \frac{1}{2}\end{bmatrix}$$

若 $p(x_1)=\frac{1}{2}$，$p(x_2)=p(x_3)=\frac{1}{4}$，试求最佳译码时的平均错误概率。

5.3 考虑一个码长为 4 的二元码，其码字为 $\boldsymbol{W}_1=0000$，$\boldsymbol{W}_2=0011$，$\boldsymbol{W}_3=1100$，$\boldsymbol{W}_4=1111$。若将码字送入一个二元对称信道，该信道的单符号错误概率为 p，且 $p<0.01$，输入码字的概率分布为 $p(\boldsymbol{W}_1)=\frac{1}{2}$，$p(\boldsymbol{W}_2)=\frac{1}{8}$，$p(\boldsymbol{W}_3)=\frac{1}{8}$，$p(\boldsymbol{W}_4)=\frac{1}{4}$。试找出一种译码规则，使平均错误概率 P_E 最小。

5.4 设有一离散无记忆信道，其信道矩阵为

$$\boldsymbol{P}=\begin{bmatrix}\frac{1}{2} & \frac{1}{2} & 0 & 0 & 0\\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0\\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0\\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2}\\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2}\end{bmatrix}$$

（1）计算信道容量 C。

（2）找出一个码长为 2 的重复码，其信息传输率为 $\frac{1}{2}\log 5$。当输入码字为等概分布时，如果按最大似然译码规则设计译码器，求译码器输出端的平均错误概率。

5.5 已知一个线性分组码的生成矩阵为

$$\boldsymbol{G}=\begin{bmatrix}1 & 0 & 0 & 0 & 1 & 1 & 1\\ 0 & 1 & 0 & 0 & 1 & 0 & 1\\ 0 & 0 & 1 & 0 & 0 & 1 & 1\\ 0 & 0 & 0 & 1 & 1 & 1 & 0\end{bmatrix}$$

试求该码组的校验矩阵。

5.6 已知某系统汉明码的校验矩阵为

$$\boldsymbol{H}=\begin{bmatrix}1 & 1 & 1 & 0 & 1 & 0 & 0\\ 0 & 1 & 1 & 1 & 0 & 1 & 0\\ 1 & 1 & 0 & 1 & 0 & 0 & 1\end{bmatrix}$$

试求其生成矩阵。当输入序列为 110101101010 时，求编码器编出的码序列。

5.7 已知(6, 3)线性分组码的全部码字如下：

110100
110011
011010
011101
101001
000111
101110
000000

问该码能纠正单个错误吗？构造该码组的生成矩阵和校验矩阵。

5.8 已知(n, k)码的校验矩阵为

$$\boldsymbol{H}=\begin{bmatrix}100 & 100 & 110\\ 101 & 010 & 010\\ 011 & 100 & 001\\ 101 & 011 & 101\end{bmatrix}$$

试求信息元 k、编码效率 r 和码的最小距离。

5.9 已知(n, k)码的校验矩阵为

$$\boldsymbol{H}=\begin{bmatrix}1&0&0&1&0&0&1&1&0\\ 1&0&1&0&1&0&0&1&0\\ 0&1&1&1&0&0&0&0&1\\ 1&0&1&0&1&1&1&0&1\end{bmatrix}$$

(1) 试求 $n=?$ $k=?$ 能生成多少个码字？

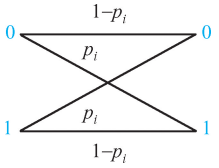
(2) 求该码的生成矩阵。

5.10 设线性分组码的校验矩阵

$$\boldsymbol{H}=\begin{bmatrix}1&0&0&1&0&0&1&1&0\\ 1&0&1&0&1&0&0&1&0\\ 0&1&1&1&0&0&0&0&1\\ 1&0&1&0&1&1&1&0&1\end{bmatrix}$$

试求该矩阵的标准型校验矩阵和生成矩阵。

5.11 考虑一个时变离散无记忆信道，如题图 5.11 所示。给定 $X_1, X_2, \cdots, X_n$ 的情况下，$Y_1, Y_2, \cdots, Y_n$ 条件独立，它们之间的条件分布 $p(\boldsymbol{y}\mid\boldsymbol{x})=\prod_{i=1}^{n}p_i(y_i\mid x_i)$。$\boldsymbol{X}=(X_1, X_2, \cdots, X_n)$，$\boldsymbol{Y}=(Y_1, Y_2, \cdots, Y_n)$。求$\max_{p(x)}\boldsymbol{I}(\boldsymbol{X};\boldsymbol{Y})$。



题图 5.11

5.12 某一信道，其输入为 X 的符号集为$\left\{0, \frac{1}{2}, 1\right\}$，输出 Y 的符号集为$\{0, 1\}$，信道矩阵为

$$\boldsymbol{P}=\begin{bmatrix}1 & 0\\ \frac{1}{2} & \frac{1}{2}\\ 0 & 1\end{bmatrix}$$

现有 4 个消息的信源通过这信道传输（消息等概率出现）。若对信源进行编码，我们选这

样一种码 $C:\left\{\left(x_1,x_2,\frac{1}{2},\frac{1}{2}\right)\right\}$，$x_i=0(i=1,2)$的码长为 $n=4$，并选取这样的译码规则

$$f(y_1,y_2,y_3,y_4)=\left(y_1,y_2,\frac{1}{2},\frac{1}{2}\right)$$

(1) 这样编码后信息传输率等于多少?

(2) 证明在选用的译码规则下，对所有码字有 $P_E=0$。

5.13 证明最小码间距离为 D_{min} 的码用于二元对称信道能够纠正小于 $D_{min}/2$ 个错误的所有组合。

5.14 证明(n, k)线性码的最小码间距离不能超过 $n-k+1$。

5.15 给定二元对称信道，其输入符号为等概分布，单个符号的错误传递概率是 0.01。求当代码组长度 $n=80$ 时，误码率 P_E 是多少?

5.16 设一离散无记忆信道的输入符号集为 $X=\{x_1,x_2,\cdots,x_r\}$，输出符号集为 $Y=\{y_1,y_2,\cdots,y_s\}$，信道转移概率为 $p(y_j\mid x_i)(i=1,2,\cdots,r;\ j=1,2,\cdots,s)$。若译码器以概率 γ_{ij} $(i=1,2,\cdots,r)$对收到的 y_j 判决为 x_i。试证明：对于给定的输入分布，任何随机判决方法得到的错误概率不低于最大后验概率译码时的错误概率。

5.17 设$(6, 3)$二元线性码的生成矩阵

$$\boldsymbol{G}=\begin{bmatrix}1&0&1&0&1&1\\0&1&1&1&1&0\\0&0&0&1&1&1\end{bmatrix}$$

(1) 试找出 $\boldsymbol{G}$ 的行缩减梯形法式表示。

(2) 求监督矩阵 $\boldsymbol{H}$。

(3) 找出最小重陪集首项。

(4) 在 B. S. C 信道中对接收矢量 111010、000011、101010 进行译码。

5.18 设有码如题表 5.18 所示。

(1) 找出生成矩阵 $\boldsymbol{G}$ 和监督矩阵 $\boldsymbol{H}$。

(2) 在 BSC 信道下给出最大似然译码的译码表。

(3) 求正确译码的概率。

题表 5.18

信　息	码　字
00	0000
01	01101
10	10111
11	11010

5.19 试证：

(1) 二元线性码中码字重必全为偶数或奇偶各半。

(2) (n, k)码的平均码字重不超过$\frac{n}{2}$。

5.20 设一离散无记忆信道转移概率矩阵为

$$\boldsymbol{P}=\begin{bmatrix}\frac{1}{2}&\frac{1}{2}&0\\0&\frac{1}{2}&\frac{1}{2}\\\frac{1}{2}&0&\frac{1}{2}\end{bmatrix}$$

对信源输出做二次重复编码。

(1) 对二次重复编码，请给出极大似然译码规则。

（2）试求（1）中的平均译码错误概率 P_E。

5.21 考虑一个二进制(5, 3)线性分组码，其生成矩阵为

$$\boldsymbol{G}=\begin{bmatrix}1&0&0&0&1\\0&1&0&1&1\\0&0&1&1&0\end{bmatrix}$$

（1）给出此码的所有码字，完成类似表 5.7 的表。

（2）请写出此码的一致校验矩阵。

（3）写出所有四个伴随式及其相应的、重量最小的四个错误图样来。

5.22 我们学过三次重复编码以及这种编码在一个错误概率为 p 的二进制对称信道上传输时的译码方法，最终的译码错误概率为 $3p^2-2p^3$，信息传输率为 $r=\frac{1}{3}$。针对四次重复编码，其信息传输率为 $r=\frac{1}{4}$。请问：

（1）其最优的译码方案是什么？

（2）其译码错误概率是多少？

（3）请将它与三次重复编码进行比较。

第 6 章　限失真信源编码

第 4 章讨论了离散信源的无失真信源编码，它是一种冗余度压缩编码，可以对信源输出的信息进行有效的表示，编码不会带来失真。从信号携带信息的角度看，编译码前后的信号具有相同的信息熵，因此冗余度压缩编码是无失真的保熵的编码。

但是无失真的保熵的编码并非是必需的，有时候不可能实现。比如，电话通信时，由于人耳接受的信号带宽和分辨率是有限的，我们可以把频谱范围 100 Hz ~ 7 kHz 的语音信号去掉低端和高端的频率，只保留 300 ~ 3400Hz 的信号。这样虽然会有一些失真，但是不会影响可懂度。再如，在传送活动图像时，由于人眼的视觉暂留特性，只需每秒传送 25 帧的静止图像，人们看到的就是连贯的活动图像了。

另一方面，由于受到信息存储、处理或传输设备的限制，而不得不对信源输出的信号进行某种近似的表示。比如，实际信源的输出常常是连续的消息，连续信源的绝对熵 $H(S)$ 是无限大的。若要求无失真地传送连续信源的消息，则信息传输率 R 也应为无限大。而在信道中，由于带宽总是有限的，所以信道容量总是受到限制，而实际信源输出的信息率总是大大超过信道容量（$R > C$），因此也就不可能实现完全无失真地传输信源的消息。如果要把连续信源的消息离散化，由于信源熵为无穷大，根据无失真信源编码定理，需要用无穷多个比特才能完全无失真地描述它，这在实际中也是做不到的，因此必然会带来一定程度的失真。所以，在实际生活中，通常总是要求在保证一定质量的前提下，信宿近似地再现信源输出的信息。因此，实际的信息传输率还可以降低。

在允许一定失真程度的条件下，怎样用尽可能少的码符号来表达信源的信息，也就是信源熵所能压缩的极限或者编码后信源输出的信息率压缩的极限值，这就是本章要讨论的问题——限失真信源编码问题或者信息率失真理论。限失真信源编码也称保真度准则下的信源编码、熵压缩编码，它是量化、数模转换、频带压缩和数据压缩的理论基础。

如果说无失真的冗余度压缩编码主要是针对离散信源，那么，有失真的熵压缩编码主要针对连续信源。本章讨论的是离散无记忆信源的限失真信源编码理论，这样便于理解限失真信源编码理论的基本概念。

我们讨论的物理模型仍然是信源编码器，编码器的输入符号集 $X = \{x_1, x_2, \cdots, x_r\}$，输出符号集 $Y = \{y_1, y_2, \cdots, y_s\}$。编码器可以视为一个广义的信道，$X$ 是信道的输入，Y 是信道的输出。与无失真信源编码不同，这时从输入到输出的映射是一个多对一的映射，是不可逆的，信源符号序列和码符号序列之间的差异就是编码时引入的失真。

6.1　失真测度

下面研究在给定允许失真的条件下，是否可以设计一种信源编码使信息传输率为最低。

为了定量的描述信息率和失真的关系，必须先规定失真的测度。

6.1.1 失真函数

设有离散无记忆信源

$$\begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \cdots & x_r \\ p(x_1) & p(x_2) & \cdots & p(x_r) \end{bmatrix}$$

经过信道传输后输出随机变量 Y 的概率空间

$$\begin{bmatrix} Y \\ P \end{bmatrix} = \begin{bmatrix} y_1 & y_2 & \cdots & y_s \\ p(y_1) & p(y_2) & \cdots & p(y_s) \end{bmatrix}$$

对于每对(x_i, y_j)，指定一个非负的函数 $d(x_i, y_j) \geqslant 0 (i = 1, 2, \cdots, r; j = 1, 2, \cdots, s)$，表示信源发出一个符号 x_i，而在接收端再现为 y_j 所引起的误差或失真的大小，称 $d(x_i, y_j)$ 为单个符号的失真度或失真函数，通常较小的 d 值代表较小的失真，而 $d(x_i, y_j) = 0$ 表示没有失真。

由于信源 X 有 r 个符号，信道输出 Y 有 s 个符号，所以 $d(x_i, y_j)$ 有 $r \times s$ 个。这 $r \times s$ 个非负的函数可以排列成矩阵形式，即

$$\boldsymbol{D} = \begin{bmatrix} d(x_1, y_1) & d(x_1, y_2) & \cdots & d(x_1, y_s) \\ d(x_2, y_1) & d(x_2, y_2) & \cdots & d(x_2, y_s) \\ \vdots & \vdots & \ddots & \vdots \\ d(x_r, y_1) & d(x_r, y_2) & \cdots & d(x_r, y_s) \end{bmatrix}$$

$\boldsymbol{D}$ 称为失真矩阵，它是一个 $r \times s$ 阶矩阵。

失真函数是根据人们的实际需要和失真引起的损失、风险大小等人为规定的。常用的失真函数如下。

1. 汉明失真

$$d(x_i, y_j) = \begin{cases} 0 & x_i = y_j \\ 1 & x_i \neq y_j \end{cases}$$

例如，在离散对称信道（$r = s$）中，我们经常定义单个符号失真度为汉明失真，表示当再现的接收符号与发送的信源符号相同时，就不存在失真和错误，所以失真度 $d(x_i, y_j) = 0$；当再现的接收符号与发送符号不同时，就有失真存在，而且，认为发送符号与再现符号不同时所引起的失真都相同，所以失真度 $d(x_i, y_j)$（$x_i \neq y_j$）为常数，常取为 1，这种失真称为汉明失真。汉明失真矩阵 $\boldsymbol{D}$ 通常为方阵，且对角线上的元素为 0，即

$$\boldsymbol{D} = \begin{bmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & 0 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 0 \end{bmatrix}$$

$\boldsymbol{D}$ 是 $r \times r$ 阶矩阵。

2. 平方误差失真函数

$$d(x_i, y_j) = (x_i - y_j)^2$$

如果 x_i、y_j 代表信源输出信号的幅度值，则上式意味着较大的幅度差值要比较小的幅度差值引起的失真更为严重，严重程度用平方表示。例如，当信道 $r=s=3$，输入 $X=\{0,1,2\}$，输出 $Y=\{0,1,2\}$ 时，平方误差失真矩阵为

$$\boldsymbol{D}=\begin{bmatrix}0 & 1 & 4\\ 1 & 0 & 1\\ 4 & 1 & 0\end{bmatrix}$$

【例 6.1】设信道输入 $X=\{0,1\}$，输出 $Y=\{0,1,2\}$，规定失真函数 $d(0,0)=d(1,1)=0$，$d(0,1)=d(1,0)=1$，$d(0,2)=d(1,2)=0.5$，求 $\boldsymbol{D}$。

【解】

$$\boldsymbol{D}=\begin{bmatrix}0 & 1 & 0.5\\ 1 & 0 & 0.5\end{bmatrix}$$

这是一个二元删除信道。

以上是单个符号的失真函数，我们可以推广得到长度为 N 的信源符号序列的失真函数。设信源输出的符号序列 $\boldsymbol{X}=X_1X_2\cdots X_N$，其中每个随机变量 $X_1,X_2,\cdots,X_N$ 均取自于同一符号集 $X=\{x_1,x_2,\cdots,x_r\}$，所以共有 r^N 个不同的信源符号序列 $\boldsymbol{x}_i$，而接收符号序列为 $\boldsymbol{Y}=Y_1Y_2\cdots Y_N$，其中每个随机变量 $Y_1,Y_2,\cdots,Y_N$ 取值于同一符号集 $Y=\{y_1,y_2,\cdots,y_s\}$，共有 s^N 个不同的接收符号序列 $\boldsymbol{y}_j$。

【定义 6-1】设发送序列为

$$\boldsymbol{x}_i=[x_{i_1}\ x_{i_2}\cdots\ x_{i_N}]$$

接收序列为

$$\boldsymbol{y}_j=[y_{j_1}\ y_{j_2}\cdots\ y_{j_N}]$$

则定义序列的失真度为

$$\begin{aligned}d(\boldsymbol{x}_i,\boldsymbol{y}_j) &= d(x_{i_1}x_{i_2}\cdots x_{i_N},y_{j_1}y_{j_2}\cdots y_{j_N})\\ &= d(x_{i_1},y_{j_1})+d(x_{i_2},y_{j_2})+\cdots+d(x_{i_N},y_{j_N})\\ &= \sum_{k=1}^{N}d(x_{i_k},y_{j_k})\end{aligned}$$

即信源序列的失真度等于序列中对应单个符号失真度之和。取不同的 $\boldsymbol{x}_i$ 和 $\boldsymbol{y}_j$，其 $d(\boldsymbol{x}_i,\boldsymbol{y}_j)$ 不同，写成矩阵形式 $\boldsymbol{D}(N)$ 时，是 $r^N\times s^N$ 阶矩阵。

【例 6.2】假设信源输出序列 $\boldsymbol{X}=X_1X_2X_3$，其中每个随机变量均取值于 $X=\{0,1\}$。经信道传输后的输出为 $\boldsymbol{Y}=Y_1Y_2Y_3$，其中每个随机变量均取值于 $Y=\{0,1\}$。定义失真函数 $d(0,0)=d(1,1)=0$，$d(0,1)=d(1,0)=1$，求失真矩阵 $\boldsymbol{D}(N)$。

【解】

由序列的失真函数的定义，可以求出

$$d(000,000)=d(0,0)+d(0,0)+d(0,0)=0$$

$$d(000,001)=d(0,0)+d(0,0)+d(0,1)=1$$

同理，可得到矩阵其他元素的数值：

$$\boldsymbol{D}(N)=\begin{bmatrix}0&1&1&2&1&2&2&3\\1&0&2&1&2&1&3&2\\1&2&0&1&2&3&1&2\\2&1&1&0&3&2&2&1\\1&2&2&3&0&1&1&2\\2&1&3&2&1&0&2&1\\2&3&1&2&1&2&0&1\\3&2&2&1&2&1&1&0\end{bmatrix}$$

6.1.2 平均失真

单个符号失真度 $d(x_i,y_j)$ 随 x_i、y_j 的不同而不同，即对于不同的信源符号和不同的接收符号，$d(x_i,y_j)$ 是不同的。为了从总体上描述系统的失真情况，定义信源的平均失真度为

$$\begin{aligned}\overline{D}&=E[d(x_i,y_j)]\\&=\sum_{i=1}^{r}\sum_{j=1}^{s}p(x_iy_j)d(x_i,y_j)\\&=\sum_{i=1}^{r}\sum_{j=1}^{s}p(x_i)p(y_j\mid x_i)d(x_i,y_j)\end{aligned}$$

它是在 XY 的联合概率空间求平均。平均失真度已对信源和信道进行了统计平均，所以此值描述了某一信源在某一信道下的失真大小。

对于 N 维信源符号序列的平均失真度为

$$\begin{aligned}\overline{D}(N)&=E[d(\boldsymbol{x}_i,\boldsymbol{y}_j)]\\&=\sum_{i=1}^{r^N}\sum_{j=1}^{s^N}p(\boldsymbol{x}_i)p(\boldsymbol{y}_j\mid\boldsymbol{x}_i)d(\boldsymbol{x}_i,\boldsymbol{y}_j)\\&=\sum_{i=1}^{r^N}\sum_{j=1}^{s^N}p(\boldsymbol{x}_i)p(\boldsymbol{y}_j\mid\boldsymbol{x}_i)\sum_{k=1}^{N}d(x_{i_k},y_{j_k})\end{aligned}$$

当信源与信道都是无记忆时，由概率关系可得

$$\overline{D}(N)=\sum_{k=1}^{N}\overline{D}_k$$

其中，$\overline{D}_k$ 是指信源序列第 k 个分量的平均失真度，而信源单个符号的的平均失真度为

$$\begin{aligned}\overline{D}_N&=\frac{1}{N}\overline{D}(N)\\&=\frac{1}{N}\sum_{i=1}^{r^N}\sum_{j=1}^{s^N}p(\boldsymbol{x}_i)p(\boldsymbol{y}_j\mid\boldsymbol{x}_i)d(\boldsymbol{x}_i,\boldsymbol{y}_j)\end{aligned}$$

当信源与信道无记忆时，有

$$\begin{aligned}\overline{D}_N&=\frac{1}{N}\overline{D}(N)\\&=\frac{1}{N}\sum_{k=1}^{N}\overline{D}_k\end{aligned}$$

注意，$\overline{D}(N)$、$\overline{D}_N$ 和$\overline{D}_k$ 表示的意义不同。

如果离散信源是平稳信源，即 $p(x_{i_k})=p(x_i)$，信道又是平稳信道，即

$$p(y_{j_k}\mid x_{i_k})=p(y_j\mid x_i)\qquad k=1,2,\cdots N$$

则 $\overline{D}_k=\overline{D}$，$\overline{D}(N)=N\overline{D}$，即离散无记忆平稳信源通过离散无记忆平稳信道，其信源序列的平均失真度等于单个符号平均失真度的 N 倍，并且 $\overline{D}_N=\overline{D}$。

6.2 信息率失真函数

信息率失真理论研究在给定的允许失真的条件下，设计一种信源编码，使信息传输率为最低，这个最低的信息传输率称为信息率失真函数。为了方便起见，我们把编码器视为一个信道，经过限失真信源编码后产生的失真视为经过一个有噪信道产生的失真。

6.2.1 D 失真许可信道

如果要求信源编码后的平均失真度 $\overline{D}$ 小于允许的失真 D，即 $\overline{D}\leqslant D$，称为保真度准则。N 维信源序列的保真度准则是 $\overline{D}(N)\leqslant ND$。

平均失真度 $\overline{D}$ 不仅与单个符号的失真度有关，还与信源的概率分布与信道的转移概率有关。当信源和单个符号失真度固定，即 $P(X)$ 和 $d(x_i,y_j)$ 给定时，选择不同的信道相当于选择不同的编码方法，所得的平均失真度 $\overline{D}$ 不同，有些信道 $\overline{D}\leqslant D$，些信道 $\overline{D}>D$。凡满足保真度准则 $\overline{D}\leqslant D$ 的信道称为 D 失真许可的试验信道。所有 D 失真许可的试验信道的集合用 B_D 表示，即

$$B_D=\{p(y_j\mid x_i):\overline{D}\leqslant D\}\qquad i=1,2,\cdots,r;\ j=1,2,\cdots,s$$

对于离散无记忆信源的 N 次扩展信源和离散无记忆信道的 N 次扩展信道，相应的 D 失真许可的试验信道为

$$B_{D(N)}=\{p(\boldsymbol{y}_j\mid \boldsymbol{x}_i):\overline{D}(N)\leqslant ND\}\qquad i=1,2,\cdots,r^N;\ j=1,2,\cdots,s^N$$

6.2.2 信息率失真函数的定义

如果信源输出的信息率为 R，在信道容量为 C 的信道上传输，如果 $R>C$，就会引起失真，我们需要对信源进行压缩，使压缩后信源输出的信息率 R^* 小于信道容量 C。压缩的过程中会引入失真，但我们可以控制失真在一个可控的范围内，即满足保真度准则。从另一方面来说，我们总希望在满足保真度准则以后，压缩后的信息传输率 R^* 尽可能地小。这个 R^* 可以用平均互信息 $I(X;Y)$ 来表示，压缩过程中引入的失真可以用 $H(X\mid Y)$ 表示。

我们的任务是，在满足保真度准则的 D 失真许可的试验信道集合 B_D 中寻找一个信道 $p(y_j\mid x_i)$，使它的 $I(X;Y)$ 达到最小。这个 $I(X;Y)$ 是允许失真 D 的函数，记为 $R(D)$，即

$$R(D)=\min_{p(y_j\mid x_i)\in B_D} I(X;Y)$$

$R(D)$ 就是信息率失真函数，也称为率失真函数。根据计算时对数的底不同，它的单位可以是比特/信源符号、哈特莱/信源符号或奈特/信源符号。

对于 N 维信源符号序列，同样可以得其信息率失真函数：

$$R_N(D)=\min_{p(\boldsymbol{y}_j\mid\boldsymbol{x}_i)\in B_{D(N)}} I(\boldsymbol{X};\boldsymbol{Y})$$

当信源和信道均为无记忆时，$I(\boldsymbol{X};\boldsymbol{Y})=NI(X;Y)$，所以

$$R_N(D)=NR(D)$$

它是在所有满足平均失真度$\overline{D}(N)\leqslant ND$ 的 N 维试验信道集合中，寻找某个信道使 $I(X;Y)$ 取极小值。因为平均失真度$\overline{D}(N)$ 与长度 N 有关，所以，在其他条件（信源概率分布，单个符号的失真度）相同的条件下，对于不同的 N，$R_N(D)$ 是不同的。

对于给定的信源，在满足保真度准则$\overline{D}\leqslant D$ 的前提下，信息率失真函数 $R(D)$ 是信源输出的信息率允许压缩到的最小值。因为 $I(X;Y)$ 是 $p(y_j\mid x_i)$ 的下凸函数，所以在 B_D 集合中一定存在一个 $I(X;Y)$ 的最小值。

从数学上来看，平均互信息 $I(X;Y)$ 是信源概率分布 $p(x_i)$ 的上凸函数，又是信道传递概率 $p(y_j\mid x_i)$ 的下凸函数，因此信道容量 C 和信息率失真函数 $R(D)$ 具有对偶性。

信道容量 $C=\max_{p(x_i)} I(X;Y)$ 是指在信道固定前提下，选择一种信源概率分布使信息传输率最大（求极大值），反映了信道传输信息的能力，是信道可靠传输的最大信息传输率。信道容量与信源无关，是信道特性的参量，不同的信道其信道容量不同。

信息率失真函数 $R(D)=\min_{p(y_j\mid x_i):\overline{D}\leqslant D} I(X;Y)$ 是对于给定的信源，在满足保真度准则的条件下的信息传输率的最小值，反映了满足一定失真度的条件下信源可以压缩的程度，即满足失真要求而再现信源消息所必须获得的最少平均信息量。$R(D)$ 是信源特性的参量，$R(D)$ 一旦求到就与求极值过程中选择的试验信道无关，不同的信源 $R(D)$ 不同。这两个概念的适用范围不同。研究信道容量 C 是为了解决在已知信道中尽可能多地传输信息量的问题，是为了充分利用已给信道，使传输的信息量最大而错误概率任意小，这是信道编码的问题。

研究信息率失真函数是为了解决在已知信源和允许失真度条件下，使信源输出的信息率尽可能小，也就是在允许一定失真度 D 的条件下，使信源必须传输给信宿的信息量最少，尽可能用最少的码符号来传输信源信息，使信源的信息可以尽快传输出去，以提高通信的有效性。这是信源编码问题。

6.2.3 信息率失真函数 $R(D)$ 的性质

1. $R(D)$ 的定义域

D 是允许的平均失真度，$R(D)$ 是对应于 D 的一个确定的信息传输率。对于给定的信源，允许失真 D 不同，$R(D)$ 就不同，它是允许失真度 D 的函数。求 $R(D)$ 的定义域，即 D 的取值范围，是在信源的概率分布和失真函数 $d(x_i,y_j)$ 给定的情况下，在不同的试验信道下（即 $p(y_j\mid x_i)$ 不同），求得$\overline{D}$的可能的取值范围。平均失真度$\overline{D}$是非负实函数 $d(x_i,y_j)$ 的数学期望，因此$\overline{D}$也是一个非负的实数。若$\overline{D}$的下限是 0，那么允许失真度 D 的下限也必然是 0，这是不允许任何失真的情况。平均失真度$\overline{D}$能否达到下限值 0，与单个符号的失真函数的定

义有关。

$$D_{\min} = \min \sum_i \sum_j p(x_i)p(y_j \mid x_i)d(x_i,y_j)$$

$$= \sum_i p(x_i) \min \sum_j p(y_j \mid x_i)d(x_i,y_j)$$

可以这样选择试验信道，即选择转移概率：对于每个 x_i，找出一个使 $d(x_i,y_j)$ 最小的 y_j，令其 $p(y_j \mid x_i)=1$，而其他转移概率为 0。这样可以得到

$$D_{\min} = \sum_i p(x_i) \min_j d(x_i,y_j)$$

只有当失真矩阵每一行至少有一个零元素时，信源的平均失真度才能达到下限值 0，否则 $D_{\min}>0$。在实际情况中，一般 $D_{\min}=0$，它表示信源不允许任何失真存在，直观地理解，这时的信息率至少应等于信源输出的平均信息量——信源熵，即 $R(0)=H(X)$。对于连续信源，$R(0)=H(X)\to\infty$。

当失真矩阵除了满足 $D_{\min}=0$ 的条件，即每行至少有一个零以外，某些列还有不止一个 0 时，说明信源符号集有些符号可以压缩、合并而不带来任何失真，压缩后的信息率必然减小，这时 $R(0)$ 可以小于 $H(X)$。

【例 6.3】删除信道 $X=\{0,1\}$，$Y=\{0,1,2\}$，失真矩阵 $\boldsymbol{D}=\begin{bmatrix}0 & 1 & \frac{1}{2}\\ 1 & 0 & \frac{1}{2}\end{bmatrix}$，求 $D_{\min}$。

【解】最小允许失真度为

$$D_{\min} = \sum_{i=1}^{2} p(x_i) \min_j d(x_i,y_j)$$

$$= \sum_{i=1}^{2} p(x_i) \cdot 0 = 0$$

$D_{\min}=0$ 时，不管何种信源分布都能达到最小允许失真度。满足这个最小允许失真度的试验信道是一个无噪无损的试验信道 $\boldsymbol{P}=\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\end{bmatrix}$，且 B_D 中只有这样一个信道。这时

$$I(X;Y)=H(X)$$

$$R(0) = \min_{p(y_j \mid x_i)\in B_D} I(X;Y) = H(X)$$

【例 6.4】设信源 $\begin{bmatrix}X\\ P(X)\end{bmatrix}=\begin{bmatrix}0 & 1 & 2\\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3}\end{bmatrix}$，信宿 $Y=\{0,1\}$，失真矩阵 $\boldsymbol{D}=\begin{bmatrix}0 & 1\\ \frac{1}{2} & \frac{1}{2}\\ 1 & 0\end{bmatrix}$，求 $D_{\min}$。

【解】

$$D_{\min}=\frac{1}{3}\cdot 0+\frac{1}{3}\cdot\frac{1}{2}+\frac{1}{3}\cdot 0=\frac{1}{6}$$

满足这个最小允许失真度的试验信道是

$$\begin{cases} p(y_1 \mid x_1) = 1 \\ p(y_2 \mid x_1) = 0 \\ p(y_1 \mid x_2) + p(y_2 \mid x_2) = 1 \\ p(y_1 \mid x_3) = 0 \\ p(y_2 \mid x_3) = 1 \end{cases}$$

$B_{D_{\min}}$的试验信道有无穷多个，因为$p(y_1 \mid x_2)$和$p(y_2 \mid x_2)$可以为多种组合，只要满足均大于等于0且和为1就行。因为输出可能对应多个输入，因此信道疑义度$H(X \mid Y) \neq 0$，所以

$$R(D_{\min}) = R\left(\frac{1}{6}\right) = \min_{p(y_j \mid x_i) \in B_{D\min}} I(X;Y) < H(X)$$

平均失真度也有最大值$D_{\max}$。根据率失真函数的定义，$R(D)$是在一定约束条件下平均互信息的$I(X;Y)$的极小值。由于$I(X;Y)$是非负的，$R(D)$也必然是非负的，其下限值必为0。从直观上理解，不允许任何失真时，平均传送一个信源符号所需的信息率最大，即必须等于信源熵，这就是平均互信息的上限值。当允许一定的失真存在时，传送信源符号所需的信息率就小些。反过来说，信息率越小，失真就越大，当$R(D)$等于0时，对应的平均失真最大，这就是$R(D)$函数定义域的上限值$D_{\max}$。

事实上，满足$R(D) = 0$的D可以有无穷多个，取最小的一个定义为$D_{\max}$。当$D \geqslant D_{\max}$时，$R(D) = 0$。

当$R(D) = 0$，这个最小的$I(X;Y) = 0$，这时相当于X和Y统计独立的情况。这意味着在接收端收不到信源发送的任何信息，与信源不发送任何信息是等效的。换句话说，传送信源符号的信息率可以压缩为0。

当$D \geqslant D_{\max}$时，$R(D) = 0$。只要X和Y相互独立的试验信道就可以使得$R(D) = 0$，这时$p(y_j \mid x_i) = p(y_j)$。这样，不同的$p(y_j)$都可以使得$R(D) = 0$，但是所造成的$\overline{D}$有不同值，选取其中的最小值定义为$D_{\max}$，即

$$D_{\max} = \min \sum_j p(y_j) \sum_i p(x_i) d(x_i, y_j)$$

由于信源概率分布$p(x_i)$和失真函数$d(x_i, y_j)$已经给定，因此求$D_{\max}$相当于寻找一种信道输出分布$p(y_j)$，使上式右端最小。如果选取$\sum_i p(x_i) d(x_i, y_j)$的最小值对应的$p(y_j) = 1$，而令其他$\sum_i p(x_i) d(x_i, y_j)$对应的$p(y_j) = 0$，则

$$D_{\max} = \min_j \sum_i p(x_i) d(x_i, y_j)$$

【例6.5】二元信源$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \\ 0.4 & 0.6 \end{bmatrix}$，$\boldsymbol{D} = \begin{bmatrix} \alpha & 0 \\ 0 & \alpha \end{bmatrix}$，计算$D_{\max}$。

【解】

$$D_{\max} = \min(0.4\alpha, 0.6\alpha) = 0.4\alpha$$

综上所述，率失真函数$R(D)$的定义域为$(D_{\min}, D_{\max})$，一般情况下$D_{\min} = 0$，$R(D) = H(X)$，$R(D_{\max}) = 0$。而$D_{\min} < D < D_{\max}$时，$H(X) > R(D) > 0$。

2. $R(D)$ 是关于 D 的下凸函数

【定理 6-1】$R(D)$ 是关于 D 的下凸函数，即对于任意 $0 \leqslant \alpha \leqslant 1$ 和 $D_1, D_2 \leqslant D_{\max}$，有

$$R[\alpha D_1 + (1-\alpha)D_2] \leqslant \alpha R(D_1) + (1-\alpha)R(D_2)$$

【证明】

设给定信源 X 和失真函数 $d(x_i, y_j)(i=1,2,\cdots,r;\ j=1,2,\cdots s)$，在 $R(D)$ 函数的定义域内选取两个允许失真度 D_1 和 D_2，并设两个试验信道 $p_1(y_j \mid x_i)$ 和 $p_2(y_j \mid x_i)$ 分别满足保真度准则，并达到相应的信息率失真函数 $R(D_1)$ 和 $R(D_2)$，即

$$\overline{D_1} = \sum_i \sum_j p(x_i) p_1(y_j \mid x_i) d(x_i, y_j) \leqslant D_1$$

$$\overline{D_2} = \sum_i \sum_j p(x_i) p_2(y_j \mid x_i) d(x_i, y_j) \leqslant D_2$$

并且

$$I[p_1(y_j \mid x_i)] = R(D_1)$$

$$I[p_2(y_j \mid x_i)] = R(D_2)$$

另设 $p(y_j \mid x_i) = \alpha p_1(y_j \mid x_i) + (1-\alpha) p_2(y_j \mid x_i)$，则

$$\begin{aligned}\overline{D} &= \sum_i \sum_j p(x_i) p(y_j \mid x_i) d(x_i, y_j) \\ &= \alpha \sum_i \sum_j p(x_i) p_1(y_j \mid x_i) d(x_i, y_j) + (1-\alpha) \sum_i \sum_j p(x_i) p_2(y_j \mid x_i) d(x_i, y_j) \\ &\leqslant \alpha D_1 + (1-\alpha) D_2\end{aligned}$$

所以，$p(y_j \mid x_i)$ 是满足保真度准则 $\overline{D} \leqslant \alpha D_1 + (1-\alpha)D_2$ 的试验信道。

根据率失真函数的定义，有

$$I[p(y_j \mid x_i)] \geqslant R[\alpha D_1 + (1-\alpha)D_2]$$

对于固定信源 X 来说，平均互信息是信道传递概率 $p(y_j \mid x_i)$ 的下凸函数，所以

$$\begin{aligned}I[p(y_j \mid x_i)] &\leqslant \alpha I[p_1(y_j \mid x_i)] + (1-\alpha) I[p_2(y_j \mid x_i)] \\ &= \alpha R(D_1) + (1-\alpha) R(D_2)\end{aligned}$$

综合上面两式，有

$$R[\alpha D_1 + (1-\alpha)D_2] \leqslant \alpha R(D_1) + (1-\alpha)R(D_2)$$

即率失真函数 $R(D)$ 在定义域内是允许失真度 D 的下凸函数。

证毕。

3. $R(D)$ 在定义域内是严格递减函数

$R(D)$ 具有凸状性，这意味着它在定义域内是连续的。而且，$R(D)$ 在定义域内是递减的，因为允许的失真越大，需要的信息率可以越小，根据 $R(D)$ 的定义，它是在平均失真度小于或等于允许失真度 D 的所有信道集合 B_D 中，取 $I(X;Y)$ 的最小值。允许失真度 D 扩大，那么 B_D 的集合也扩大，在扩大的 B_D 集合中找 $I(X;Y)$ 的最小值，那么结果或者不变或者比原来的小，因此 $R(D)$ 一定是递减的，即在 $0 < D < D_{\max}$ 范围内，若 $D_1 < D_2$，则 $R(D_1) \geqslant R(D_2)$。

【定理 6-2】$R(D)$ 是严格递减的，即 $R(D_1) > R(D_2)$。

【证明】如果 $R(D_1) \geqslant R(D_2)$ 中的等号成立，则在 (D_1, D_2) 中的 $R(D)$ 为常数。下面证

明(D_1,D_2)中的$R(D)$不为常数。

假设$0<D_1<D_2<D_{\max}$，$p_1(y_j\mid x_i)$和$p_m(y_j\mid x_i)$是分别达到相应的信息率失真函数$R(D_1)$和$R(D_{\max})$的两个试验信道，即

$$\overline{D_1}=\sum_i\sum_j p(x_i)p_1(y_j\mid x_i)d(x_i,y_j)\leqslant D_1$$

$$\overline{D_m}=\sum_i\sum_j p(x_i)p_m(y_j\mid x_i)d(x_i,y_j)\leqslant D_{\max}$$

且

$$I[p_1(y_j\mid x_i)]=R(D_1)$$

$$I[p_m(y_j\mid x_i)]=R(D_{\max})=0$$

那么，总能找到足够小的$\alpha>0$，满足

$$D_1<\alpha D_{\max}+(1-\alpha)D_1<D_2$$

因为$D_1+\alpha(D_{\max}-D_1)>D_1$，容易看出左边不等式成立。

而总能找到一个足够小的的α，使得

$$\alpha(D_{\max}-D_1)<D_2-D_1$$

因此，右边不等式也成立。

令$D_0=\alpha D_{\max}+(1-\alpha)D_1$，则

$$D_1<D_0<D_2$$

现在定义一个新的试验信道，设其信道传递概率为

$$p(y_j\mid x_i)=\alpha p_m(y_j\mid x_i)+(1-\alpha)p_1(y_j\mid x_i)$$

对应的

$$\begin{aligned}\overline{D}&=\sum_i\sum_j p(x_i)p(y_j\mid x_i)d(x_i,y_j)\\&=\alpha\sum_i\sum_j p(x_i)p_m(y_j\mid x_i)d(x_i,y_j)+(1-\alpha)\sum_i\sum_j p(x_i)p_1(y_j\mid x_i)d(x_i,y_j)\\&=\alpha\overline{D_m}+(1-\alpha)\overline{D_1}\\&\leqslant\alpha D_{\max}+(1-\alpha)D_1\\&=D_0\end{aligned}$$

可见，新试验信道满足保真度准则$\overline{D}\leqslant D_0$。所以

$$I[p(y_j\mid x_i)]\geqslant R(D_0)$$

由于平均互信息是信道传递概率$p(y_j\mid x_i)$的下凸函数，所以

$$\begin{aligned}I[p(y_j\mid x_i)]&\leqslant\alpha I[p_m(y_j\mid x_i)]+(1-\alpha)[p_1(y_j\mid x_i)]\\&=(1-\alpha)R(D_1)<R(D_1)\end{aligned}$$

所以，$R(D_0)<R(D_1)$。而$D_1<D_0<D_2$，所以在(D_1,D_2)区间内的$R(D)$不为常数，即$R(D_1)\geqslant R(D_2)$中的等号不成立，$R(D)$是定义域内严格递减函数。

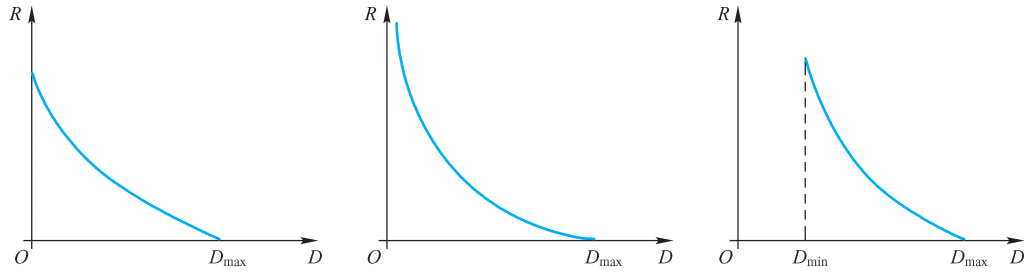
证毕。

由于信息率失真函数$R(D)$是严格的单调递减函数，因此在B_D中最小的$I(X;Y)$对应的试验信道$p(y_j\mid x_i)$必在B_D的边界上，即

$$\overline{D}=\sum_i\sum_j p(x_i)p(y_j\mid x_i)d(x_i,y_j)=D$$

所以，通常选择在$\overline{D}=D$的条件下来计算信息率失真函数$R(D)$。

根据以上性质可以画出率失真函数$R(D)$的曲线图。$R(0)=H(X)$，$R(D_{max})=0$决定了曲线的两个端点，在0和D_{max}之间$R(D)$是单调递减的下凸函数，如图6.1(a)所示。在连续信源的情况下，$R(0)\to\infty$，曲线不与$R(D)$轴相交，如图6.1(b)所示。如果$D_{min}\neq 0$，可以得到图6.1(c)。



（a）D_{min}=0时离散信源　（b）D_{min}=0时连续信源　（c）$D_{min}\neq 0$时离散信源

图6.1　信息率失真函数

6.3　限失真信源编码定理

对于无失真信源编码来说，每个信源符号（或符号序列）必须有一个对应的码字（或码字序列），信源熵不能损失，而在允许一定失真的情况下，有可能是多个信源符号（符号序列）对应一个码字（码字序列），信源输出信息率最少可减少到信息率失真函数$R(D)$。限失真信源编码定理就是关于信息率和失真关系的一个极限定理，也称为香农第三定理或保真度准则下的离散信源编码定理。

【定理6-3】设$R(D)$是离散无记忆信源的信息率失真函数并且失真函数为有限值。对于任意允许失真度$D\geqslant 0$，当码长N足够长时，一定存在一种编码，其编码后的实际传输信息率$R>R(D)$，而平均失真度$\overline{D}<D$；不存在信息传输率$R<R(D)$而平均失真度$\overline{D}<D$的任何信源编码。

限失真信源编码定理也是一个极限存在定理。它告诉我们，信息率失真函数$R(D)$是一个界限，只要实际信息传输率R大于这个界限，就可以通过信源编码技术将译码失真限制在给定的范围内，也就是说在通信的过程中虽然有失真，但仍能满足我们的要求。如果实际信息传输率$R<R(D)$，则无法满足我们的失真要求。限失真信源编码的目的是找到与信息率失真函数$R(D)$相匹配的编码，即希望R逼近$R(D)$，而无失真信源编码是寻求与信源信息熵相匹配的编码，即希望R达到$H_r(S)=R(D=0)$。

常用的限失真信源编码有量化编码、预测编码和变换编码将在6.5节介绍。

6.4　信息率失真函数的计算*

已知信源的概率分布和失真函数，就可以确定信源的信息率失真函数$R(D)$，是在约束

条件即保真度准则下求 $I(X,Y)$ 的极小值的问题。应用拉格朗日乘子法，原则上可以求出解，但是要得到明显的解析表达式是比较困难的，通常只能用参量形式来表示，或采用迭代算法用计算机求解。

6.4.1 应用参量表示式计算 $R(D)$

下面采用拉格朗日乘子法求解 $R(D)$。约束条件如下：

$$p(y_j \mid x_i) \geqslant 0 \qquad i=1,2,\cdots,n;\ j=1,2,\cdots,m$$

$$\sum_j p(y_j \mid x_i) = 1 \qquad i=1,2,\cdots,n$$

$$\sum_{ij} p(x_i)p(y_j \mid x_i)d_{ij} = D$$

求

$$I(X;Y) = \sum_{ij} p(x_i)p(y_j \mid x_i)\ln\frac{p(y_j \mid x_i)}{p(y_j)}$$

的极值。

因为 $I(X;Y)$ 是关于 $p(y_j \mid x_i)$ 的下凸函数，所以得到的极值是极小值。

设辅助函数如下：

$$F = I(X;Y) - \mu_i \sum_j p(y_j \mid x_i) - S\sum_{ij} p(x_i)p(y_j \mid x_i)d_{ij}$$

为方便起见，令 $\mu_i = p(x_i)\ln\lambda_i$，则

$$F = \sum_{ij} p(x_i)p(y_j \mid x_i)\ln\frac{p(y_j \mid x_i)}{p(y_j)} - p(x_i)\ln\lambda_i \sum_j p(y_j \mid x_i) - \sum_{ij} p(x_i)p(y_j \mid x_i)Sd_{ij}$$

令 $\dfrac{\partial F}{\partial p(y_j \mid x_i)} = 0$，得

$$\ln p(y_j \mid x_i) - \ln p(y_j) - \ln\lambda_i - Sd_{ij} = 0$$

所以

$$p(y_j \mid x_i) = \lambda_i p(y_j)e^{Sd_{ij}} \tag{6.1}$$

将式(6.1)两边同乘以 $p(x_i)$，并对 X 求和：

$$\sum_i p(x_i)p(y_j \mid x_i) = \sum_i p(x_i)\lambda_i p(y_j)\mathrm{e}^{Sd_{ij}}$$

则

$$p(y_j) = p(y_j)\sum_i p(x_i)\lambda_i \mathrm{e}^{Sd_{ij}}$$

所以

$$\sum_i p(x_i)\lambda_i \mathrm{e}^{Sd_{ij}} = 1 \tag{6.2}$$

这是 m 个方程组成的方程组，有 $n+1$ 个变量，分别为 $\lambda_i(i=1,2,\cdots,n)$ 和 S。

通过这个方程组可以求出 λ_i，S 作为参数。将式(6.1)两边对 Y 求和：

$$\sum_j p(y_j \mid x_i) = \lambda_i \sum_j p(y_j)\mathrm{e}^{Sd_{ij}} = 1$$

则得到

$$\sum_j p(y_j)\mathrm{e}^{Sd_{ij}} = \frac{1}{\lambda_i} \tag{6.3}$$

这是 n 个方程组成的方程组，有 $m+1$ 个变量，分别为 $p(y_j)(j=1,2,\cdots,m)$ 和 S，通过解这个方程组，可以将 $p(y_j)$ 用 S 为参数表示。所以

$$\begin{aligned} D &= \sum_{ij} p(x_i)p(y_j \mid x_i)d_{ij} \\ &= \sum_{ij} p(x_i)\lambda_i p(y_j)\mathrm{e}^{Sd_{ij}}d_{ij} \end{aligned} \tag{6.4}$$

通过式(6.2)和式(6.3)，可以分别确定 λ_i 和 $p(y_j)$，代入

$$I(X;Y) = \sum_{ij} p(x_i)p(y_j \mid x_i)\ln\frac{p(y_j \mid x_i)}{p(y_j)}$$

此时 $I(X;Y)$ 是关于 $p(y_j \mid x_i)$ 的一个最小值，就是要求的率失真函数，以 S 为参数表示：

$$\begin{aligned} R(S) &= \sum_{ij} p(x_i)p(y_j \mid x_i)\ln\frac{p(y_j \mid x_i)}{p(y_j)} \\ &= \sum_{ij} p(x_i)p(y_j \mid x_i)\ln\frac{\lambda_i p(y_j)\mathrm{e}^{Sd_{ij}}}{p(y_j)} \\ &= \sum_{ij} p(x_i)p(y_j \mid x_i)[\ln\lambda_i + Sd_{ij}] \\ &= \sum_{ij} p(x_i)p(y_j \mid x_i)\ln\lambda_i + \sum_{ij} p(x_i)p(y_j \mid x_i)Sd_{ij} \\ &= \sum_{ij} p(x_i)p(y_j \mid x_i)\ln\lambda_i + SD \\ &= \sum_{i} p(x_i)\ln\lambda_i + SD \end{aligned} \tag{6.5}$$

在求率失真函数时，一般要先给定信源的允许失真 D。利用式(6.4)，由 D 来确定对应的 S，则待定系数 S 可以用 D 表示。把式(6.5)中的 λ_i 和 S 都用 D 表示，就得到了 $R(D)$。

【例 6.6】二元信源的信息率失真函数。

信源

$$\begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ p & 1-p \end{bmatrix} \qquad p \leqslant \frac{1}{2}$$

输出符号集为(0，1)，失真函数定义为

$$d_{ij} = \begin{cases} 0 & i=j \\ 1 & i \neq j \end{cases} \quad (i,j=1,2)$$

求 $R(D)$。

【解】

(1) 由式(6.2)计算 λ_1 和 λ_2。

记 $p_1=p(0)=p$，$p_2=p(1)=1-p$，则

$$\begin{cases} \lambda_1 p_1 \mathrm{e}^{Sd_{11}} + \lambda_2 p_2 \mathrm{e}^{Sd_{21}} = 1 \\ \lambda_1 p_1 \mathrm{e}^{Sd_{12}} + \lambda_2 p_2 \mathrm{e}^{Sd_{22}} = 1 \end{cases}$$

把已知量代入，则

$$\begin{cases}\lambda_1 p+\lambda_2(1-p)\mathrm{e}^{S}=1\\ \lambda_1 p\mathrm{e}^{S}+\lambda_2(1-p)=1\end{cases}$$

可得

$$\lambda_1=\frac{1}{p(1+\mathrm{e}^{S})}$$

$$\lambda_2=\frac{1}{(1-p)(1+\mathrm{e}^{S})}$$

（2）由式（6.3）计算 $p(y_1)$ 和 $p(y_2)$。

$$\begin{cases}p(y_1)\mathrm{e}^{Sd_{11}}+p(y_2)\mathrm{e}^{Sd_{12}}=\dfrac{1}{\lambda_1}\\ p(y_1)\mathrm{e}^{Sd_{21}}+p(y_2)\mathrm{e}^{Sd_{22}}=\dfrac{1}{\lambda_2}\end{cases}$$

求出 $p(y_1)$ 和 $p(y_2)$，并将 λ_1、λ_2 用 S 表示，即

$$p(y_1)=\frac{\dfrac{1}{\lambda_1}\mathrm{e}^{Sd_{22}}-\dfrac{1}{\lambda_2}\mathrm{e}^{Sd_{12}}}{\mathrm{e}^{Sd_{11}+Sd_{22}}-\mathrm{e}^{Sd_{12}+Sd_{21}}}$$

$$=\frac{p-(1-p)\mathrm{e}^{S}}{1-\mathrm{e}^{S}}$$

$$p(y_2)=\frac{\dfrac{1}{\lambda_2}\mathrm{e}^{Sd_{11}}-\dfrac{1}{\lambda_1}\mathrm{e}^{Sd_{21}}}{\mathrm{e}^{Sd_{11}+Sd_{22}}-\mathrm{e}^{Sd_{12}+Sd_{21}}}$$

$$=\frac{1-p-p\mathrm{e}^{S}}{1-\mathrm{e}^{S}}$$

（3）将求得的 λ_1、λ_2 和 $p(y_1)$、$p(y_2)$ 代入式（6.4），得到平均失真度

$$D(S)=\lambda_1 p_1 p(y_1)d_{11}\mathrm{e}^{sd_{11}}+\lambda_1 p_1 p(y_2)d_{12}\mathrm{e}^{Sd_{12}}+\lambda_2 p_2 p(y_1)d_{21}\mathrm{e}^{Sd_{21}}+\lambda_2 p_2 p(y_2)d_{22}\mathrm{e}^{Sd_{22}}$$

$$=\frac{\mathrm{e}^{S}}{1+\mathrm{e}^{S}}$$

解出参量 S

$$S=\log\frac{D}{1-D}$$

（4）将参量 S 代入式（6.5），得到率失真函数

$$\begin{aligned}R(D)&=SD+p\log\lambda_1+(1-p)\log\lambda_2\\&=D\log\frac{D}{1-D}-p\log p(1+\mathrm{e}^{S})-(1-p)\log(1-p)(1+\mathrm{e}^{S})\\&=H(p,1-p)-H(D,1-D)\\&=H(p)-H(D)\end{aligned}$$

其中，第一项是信源熵，第二项是因容忍一定的失真而可以压缩的信息率。

当 $D=0$ 时，$R(D)=H(p)$。当 $D=D_{\max}=p$ 时，$R(D_{\max})=0$。

对于不同 p 值，可以得出一组 $R(D)$ 的曲线，如图 6.2 所示。可以看出，对于给定的平均失真度 D，信源分布越均匀（p 值接近 0.5），$R(D)$ 越大，可压缩性越小。反之，信源分

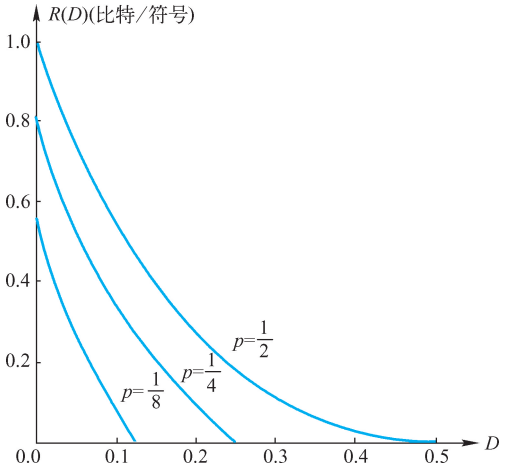


图 6.2　二元信源的信息率失真函数

布越不均匀，$R(D)$越小，可压缩性越大。因为根据最大离散熵定理，信源越趋于等概分布，其熵越大，即不确定越大，要去除这不确定性所需的信息传输率越大，而 $R(D)$正是去除信源不确定性所必需的信息传输率（在容忍一定的失真的情况下）。当 $D=D_{\max}$时，$R(D_{\max})=0$，即不用传输信息。例如，不管信源发 0 还是 1，都把它编成 1，只传输一种符号就相当于不用传输任何符号了，即 $R=0$。

【例 6.7】等概信源的信息率失真函数。

信源输出符号集 $X=\{x_1,x_2,\cdots,x_r\}$，等概分布 $p(x_i)=\dfrac{1}{r}(i=1,2,\cdots,r)$，输出符号集 $Y=\{y_1,y_2,\cdots,y_r\}$，失真函数定义为

$$d(x_i,y_j)=\begin{cases}0 & i=j\\ 1 & i\neq j\end{cases}\quad(i,j=1,2,\cdots,r)$$

求 $R(D)$。

【解】引入记号：

$$p_i=p(x_i)=\frac{1}{r}\qquad i=1,2,\cdots,r$$

$$q_j=p(y_j)\qquad j=1,2,\cdots,r$$

$$d_{ij}=d(x_i,y_j)\qquad i,j=1,2,\cdots,r$$

（1）由式(6.2)确定 λ_i

$$\begin{cases}\lambda_1+\lambda_2e^{S}+\cdots+\lambda_re^{S}=r\\ \lambda_1e^{S}+\lambda_2+\cdots+\lambda_re^{S}=r\\ \qquad\qquad\vdots\\ \lambda_1e^{s}+\lambda_2e^{s}+\cdots+\lambda_r=r\end{cases}$$

解得

$$\lambda_i=\frac{r}{1+(r-1)e^{S}}\qquad i=1,2,\cdots,r$$

（2）由式（6.3）确定 q_j

$$\begin{cases} q_1 + q_2 e^{S} + \cdots + q_r e^{S} = \dfrac{1 + (r-1)e^{S}}{r} \\ q_1 e^{S} + q_2 + \cdots + q_r e^{S} = \dfrac{1 + (r-1)e^{S}}{r} \\ \vdots \\ q_1 e^{S} + q_2 e^{S} + \cdots + q_r = \dfrac{1 + (r-1)e^{S}}{r} \end{cases}$$

解得

$$p(y_j) = q_j = \frac{1}{r} \qquad j = 1, 2, \cdots, r$$

（3）将 λ_i、q_j 代入式(6.4)，则

$$D(S) = \frac{(r-1)e^{S}}{1 + (r-1)e^{S}}$$

$$S = \ln \frac{D}{(r-1)(1-D)}$$

（4）将 S 代入式(6.5)，则

$$\begin{aligned} R(D) &= SD + \sum_{i=1}^{r} p_i \ln \lambda_i \\ &= \ln r - D\ln(r-1) + D\ln D + (1-D)\ln(1-D) \\ &= \ln r - D\ln(r-1) - H(D) \end{aligned}$$

$R(D)$的定义域为 $0 \leqslant D \leqslant 1 - \dfrac{1}{r}$，值域 $0 \leqslant R(D) \leqslant \log r$。

对于不同的 r 值，可以得到一组 $R(D)$ 曲线。对于给定的允许失真度 D，r 越大，则 $R(D)$越大，信源可压缩性越小；r 越小，$R(D)$越小，信源可压缩性越大，如图 6.3 所示。

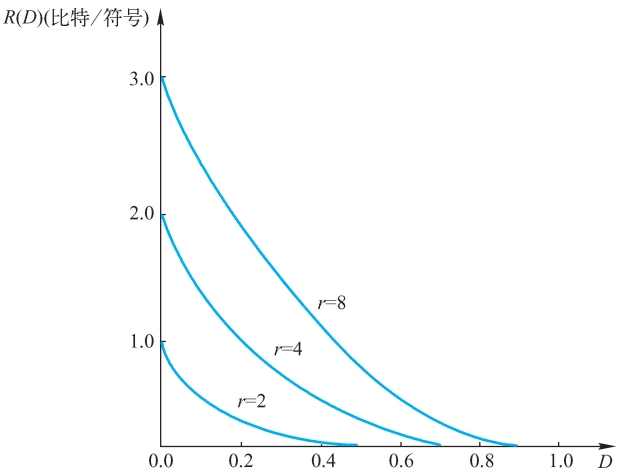


图 6.3　等概信源的信息率失真函数

由信息率失真函数可以比较编码方法的压缩效果。

【例 6.8】设信源符号集 $X = \{x_1, x_2, \cdots, x_{2r}\}$，信源概率分布为 $p(x_i) = \dfrac{1}{2r}(i = 1, 2, \cdots, 2r)$，

失真函数为

$$d(x_i,y_j)=\begin{cases}0 & x_i=y_j\\ 1 & x_i\neq y_j\end{cases}$$

求当允许的失真度为$\frac{1}{2}$时的信源输出信息率。

【解】

由信源概率分布求得信源熵

$$\begin{aligned}H(X)&=H\left(\frac{1}{2r},\frac{1}{2r},\cdots,\frac{1}{2r}\right)\\&=\log 2r\end{aligned}$$

如果对信源进行二元无失真编码，平均每个符号至少需要 $\log 2r$ 个二元码。

当允许的失真度为 0.5 时，平均每个符号需要的码元个数可以减少到什么程度呢？

假设采用如下编码方法：

① 当信源输出符号为 $x_1,x_2,\cdots,x_r$ 时，分别赋予一个码字 $y_1,y_2,\cdots,y_r$。

② 当信源输出符号为 $x_{r+1},x_{r+2},\cdots,x_{2r}$时，赋予一个相同的码字 y_r。即试验信道的输入符号集 $X=\{x_1,x_2,\cdots,x_{2r}\}$，输出符号集 $Y=\{y_1,y_2,\cdots,y_r\}$，转移概率矩阵为

$$\boldsymbol{P}=\begin{bmatrix}1&0&\cdots&0\\0&1&\cdots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&1\\0&0&\cdots&1\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&1\end{bmatrix}$$

平均失真度为

$$\begin{aligned}\overline{D}&=E[d(x_i,y_j)]\\&=\sum_{i=1}^{2r}\sum_{j=1}^{r}p(x_i)p(y_j\mid x_i)d(x_i,y_j)\\&=\sum_{i=r+1}^{2r}p(x_i)p(y_r\mid x_i)d(x_i,y_r)\\&=\sum_{i=r+1}^{2r}p(x_i)=\frac{1}{2}\end{aligned}$$

所以，上述编码方法满足保真度准则$\overline{D}\leqslant D=\frac{1}{2}$。

由于该假设信道是具有归并性能的无噪信道，故 $H(Y\mid X)=0$，从而

$$\begin{aligned}I(X;Y)&=H(Y)-H(Y\mid X)\\&=H(Y)\end{aligned}$$

由输入概率和转移概率，可求输出概率分布为

$$\begin{cases}p(y_j)=\dfrac{1}{2r} & j=1,2,\cdots,r-1\\ p(y_r)=\dfrac{r+1}{2r}\end{cases}$$

所以

$$H(Y)=H\left(\frac{1}{2r},\frac{1}{2r},\cdots,\frac{1}{2r},\frac{r+1}{2r}\right)$$

$$=\log 2r-\frac{r+1}{2r}\log(r+1)$$

即采用上述编码方案时，平均每个信源符号所需的二元码符号个数由原来的 $\log 2r$ 减少到 $\log 2r-\frac{r+1}{2r}\log(r+1)$，减少了 $\frac{r+1}{2r}\log(r+1)$ 个码元。换句话说，信源的信息率由原来的 $\log 2r$ 压缩到了 $\log 2r-\frac{r+1}{2r}\log(r+1)$，即信息率压缩了 $\frac{r+1}{2r}\log(r+1)$。这是不是达到 $R\left(\frac{1}{2}\right)$ 了呢？由例 6.7 推导的公式可以求出，$D=\frac{1}{2}$ 时，$R\left(\frac{1}{2}\right)=\log 2r-\frac{1}{2}\log(2r-1)-1$。

当 $r>1$ 时，这种编码方案的 $I(X;Y)=H(Y)>R\left(\frac{1}{2}\right)$。所以，存在更好的压缩编码方案能够进一步进行压缩，达到更好的压缩效果。

信息率失真理论给出了在给定的失真度 D 条件下，信源输出的信息率所能压缩的极限 $R(D)$，它可以作为一种尺度，衡量一种压缩编码方案的压缩效果。

6.4.2 率失真函数的迭代算法

3.2.7 节中讲过信道容量的迭代算法。这个迭代算法的关键是把平均互信息视为两个变量的函数，先固定一个变量，求平均互信息关于另一个变量的极大值，极大值求出后第二个变量作为已知量，又求平均互信息关于第一个变量的极大值，这样不断迭代最终得到符合精度要求的结果。

根据率失真函数和信道容量的对偶性可知，$I(X;Y)$ 是关于 $p(y_j\mid x_i)$ 和 $p(y_j)$ 的下凸函数，求出的极值是极小值。在给定信源分布和失真度条件下，率失真函数的迭代算法可以通过求 $I(X;Y)$ 关于 $p(y_j\mid x_i)$ 和 $p(y_j)$ 的极值来得到：

$$I(X;Y)=\sum_{ij}p(x_i)p(y_j\mid x_i)\ln\frac{p(y_j\mid x_i)}{p(y_j)}$$

$$=\sum_{ij}p(x_i)p(y_j\mid x_i)\ln p(y_j\mid x_i)-\sum_{ij}p(x_i)p(y_j\mid x_i)\ln p(y_j)$$

记为

$$I[p(y_j),p(y_j\mid x_i)]$$

（1）固定 $p(y_j\mid x_i)$，在 $\overline{D}=\sum_{ij}p(x_i)p(y_j\mid x_i)d_{ij}=D$ 和 $\sum_j p(y_j)=1$ 的约束条件下，求 $I(X;Y)$ 关于 $p(y_j)$ 的极值。

设辅助函数为

$$F=I(X;Y)-SD+\lambda\sum_j p(y_j)$$

$$=\sum_{ij}p(x_i)p(y_j\mid x_i)\ln p(y_j\mid x_i)-\sum_{ij}p(x_i)p(y_j\mid x_i)\ln p(y_j)-$$

$$S\sum_{ij}p(x_i)p(y_j\mid x_i)d_{ij}+\lambda\sum_j p(y_j)$$

$$\frac{\partial F}{\partial p(y_j)} = -\frac{\sum_i p(x_i)p(y_j \mid x_i)}{p(y_j)} + \lambda$$

令$\frac{\partial F}{\partial p(y_j)} = 0$，得

$$p(y_j) = \frac{\sum_i p(x_i)p(y_j \mid x_i)}{\lambda} \tag{6.6}$$

根据约束条件 $\sum_j p(y_j) = 1$，对式(6.6)两端求和，消去 λ，则

$$\sum_j p(y_j) = \frac{\sum_i \sum_j p(x_i)p(y_j \mid x_i)}{\lambda}$$

$$= \frac{1}{\lambda} = 1$$

所以，$\lambda = 1$。

求出极值点：

$$p(y_j)^* = \sum_i p(x_i)p(y_j \mid x_i) \tag{6.7}$$

(2) 固定 $p(y_j)$，在 $\overline{D} = \sum_{ij} p(x_i)p(y_j \mid x_i)d_{ij} = D$ 和 $\sum_j p(y_j \mid x_i) = 1(i = 1,2,\cdots,n)$ 的约束条件下，求 $I(X;Y)$ 关于 $p(y_j \mid x_i)$ 的极值。

设辅助函数为

$$Q = I(X;Y) - SD - \mu_i \sum_j p(y_j \mid x_i)$$

$$= \sum_{ij} p(x_i)p(y_j \mid x_i)\ln p(y_j \mid x_i) - \sum_{ij} p(x_i)p(y_j \mid x_i)\ln p(y_j) -$$

$$S\sum_{ij} p(x_i)p(y_j \mid x_i)d_{ij} + \mu_i \sum_j p(y_j \mid x_i)$$

为方便起见，令 $\mu_i = p(x_i)\ln\lambda_i$，则

$$Q = \sum_{ij} p(x_i)p(y_j \mid x_i)\ln p(y_j \mid x_i) - \sum_{ij} p(x_i)p(y_j \mid x_i)\ln p(y_j) -$$

$$S\sum_{ij} p(x_i)p(y_j \mid x_i)d_{ij} + p(x_i)\ln\lambda_i \sum_j p(y_j \mid x_i)$$

所以

$$\frac{\partial Q}{\partial p(y_j \mid x_i)} = p(x_i)\ln p(y_j \mid x_i) + p(x_i) - p(x_i)\ln p(y_j) - Sp(x_i)d_{ij} + p(x_i)\ln\lambda_i$$

$$= p(x_i)[\ln p(y_j \mid x_i) - \ln p(y_j) + \ln\lambda_i - sd_{ij} + 1]$$

令$\frac{\partial Q}{\partial p(y_j \mid x_i)} = 0$，得

$$\ln p(y_j \mid x_i) - \ln p(y_j) + \ln\lambda_i - Sd_{ij} + 1 = 0$$

所以

$$p(y_j \mid x_i) = p(y_j)e^{Sd_{ij}}e^{-(1+\ln\lambda_i)} \tag{6.8}$$

根据约束条件 $\sum_j p(y_j \mid x_i) = 1$，对式(6.8)两端求和，消去 $e^{-(1+\ln\lambda_i)}$，则

$$\sum_j p(y_j \mid x_i) = \sum_j p(y_j) e^{Sd_{ij}} \mathrm{e}^{-(1+\ln\lambda_i)} = 1$$

得

$$\mathrm{e}^{-(1+\ln\lambda_i)} = \frac{1}{\sum_j p(y_j)\mathrm{e}^{Sd_{ij}}}$$

极值点为

$$p(y_j \mid x_i)^* = \frac{p(y_j)\mathrm{e}^{Sd_{ij}}}{\sum_j p(y_j)\mathrm{e}^{Sd_{ij}}} \tag{6.9}$$

以上求出的 $p(y_j)^*$ 和 $p(y_j \mid x_i)^*$ 都是以 S 为参量的表达式。

（3）将 $p(y_j)^*$ 和 $p(y_j \mid x_i)^*$ 代入率失真函数的表达式，结果是以 S 为参量的表达式：

$$R(S) = \sum_{ij} p(x_i)p(y_j \mid x_i)^* \ln \frac{p(y_j \mid x_i)^*}{p(y_j)^*} \tag{6.10}$$

$$D(S) = \sum_{ij} p(x_i)p(y_j \mid x_i)^* d_{ij}$$

利用式(6.7)、式(6.9)、式(6.10)，可以对率失真函数进行迭代计算，计算步骤如下：

① 先选定一个相当大的负数作为 S 值，假定初始的信道转移概率 $p(y_j \mid x_i)^{(1)}$（$i=1,2,\cdots,n$；$j=1,2,\cdots,m$），上标表示迭代序号。

置迭代序号 $k=1$。一般假定初始的信道转移概率 $p(y_j \mid x_i)^{(1)} = \frac{1}{m}$，把假定的 $p(y_j \mid x_i)^{(1)}$ 代入式(6.7)，得到

$$p(y_j)^{(1)} = \sum_i p(x_i)p(y_j \mid x_i)^{(1)} \qquad j=1,2,\cdots,m$$

② 把得到的 $p(y_j)^{(1)}$ 代入式（6.9），则

$$p(y_j \mid x_i)^{(2)} = \frac{p(y_j)^{(1)}\mathrm{e}^{Sd_{ij}}}{\sum_j p(y_j)^{(1)}\mathrm{e}^{Sd_{ij}}} \qquad i=1,2,\cdots,n;\ j=1,2,\cdots,m$$

③ 重复上述步骤，计算出第 k 次迭代的 $p(y_j)^{(k)}$ 和 $p(y_j \mid x_i)^{(k)}$，同时利用式(6.10)计算第 k 次迭代后的 $R(S)$ 和 $D(S)$。

当 $R(S)$ 和 $D(S)$ 趋于稳定，即相邻两次迭代得到的值的差值在要求的精度范围内，则得到最终的极值。极值点的 $R(S)$ 和 $D(S)$ 就是 $R(D)-D$ 坐标图上的一个点。

④ 选定一个稍大的 S 值，重复步骤①～③，得到 $R(D)-D$ 坐标图上的第二个点。这个过程一直到率失真函数 $R(D)$ 逼近于零为止。这样就得到了 $R(D)-D$ 坐标图上的整个曲线。

与信道容量的迭代算法不同的是，信道容量的迭代算法得到的是一个具体的信道容量的数值，而率失真函数的迭代算法得到是一条曲线，是随参数 S 不同而变化的 $(R(D),D)$ 值。

6.5 常用的限失真信源编码方法

限失真信源编码方法又称为熵压缩编码方法，是不可逆的，没有对应的译码过程。常用

的熵压缩编码方法有量化编码、预测编码、变换编码、子带编码等。

6.5.1 量化编码

量化和编码是连续信源数字化的一个不可缺少的环节，而波形信源需要经过采样、量化和编码。采样就是对模拟信号时间轴的离散化，将它变成离散时间信号。量化是将连续幅度值经过量化器转变成只有有限个取值的离散幅度值。编码就是用相应的码字来表示不同的量化电平。在这三个步骤中，量化是一个多对一的处理过程，会产生量化误差，造成失真。

量化的方法有标量量化、矢量量化。标量量化是最早被研究的，每个采样使用同一个量化器进行量化，每个采样的量化都与其他所有采样无关，因此也被称为零记忆量化或一维量化。矢量量化是利用采样间的相关性进行量化，它们的量化是一次对多于一个采样值进行的，从理论上说，应该具有更好的压缩能力。但是矢量量化无论是在软件上还是在硬件上，实现起来要比标量量化复杂得多。标量置化器思想简单，且易于硬件实现，因此至今为止，仍为许多快速压缩编码系统所采用。

1. 标量量化

量化器输出 M 个电平 $\{q_1, q_2, \cdots, q_M\}$，称为量化电平。量化的过程是通过 $M-1$ 个门限电平 $T_1, T_2, \cdots, T_{M-1}$ 实现的：

$$q(x)=\begin{cases} q_1 & x \leqslant T_1 \\ q_k & T_{k-1} < x \leqslant T_k; k=2,3,\cdots,M-1 \\ q_M & x > T_{M-1} \end{cases}$$

其中，$T_1 < T_2 < \cdots < T_{M-1}$。两个门限电平之间的间隔称为量化间隔。

标量量化器主要有 3 种形式：① 均匀量化器——量化间隔是等长的，输出量化电平是在量化间隔的中点；② 非均匀量化器——量化间隔是不等长的，通常大信号采用较大的量化间隔，而小信号采用较小的量化间隔；③ 自适应量化器——量化间隔随传送数据而变。

标量量化研究的主要问题是如何使量化误差最小，使量化信噪比最大。量化误差就是实际输入值与量化值之差，反映了信号的损失情况，而量化噪声是量化误差的均方值。

2. 矢量量化

矢量量化也称为向量量化（Vector Quantization，VQ）。标量量化每次只量化一个采样值，这样的处理方法忽略了信源符号之间的相关性，因此信源的冗余度没有得到有效的压缩。而矢量量化利用相邻采样值的高度相关性，每次将 k 个采样值量化为一个编码值，这 k 个采样值构成一个 k 维的矢量。编码值被称为码字或码矢，码字的集合被称为码本或码书。

编码前，需要先生成一个码书；编码时，每个矢量与码书中的每个码字进行比较，求出相应的失真，然后用失真最小的码字的序号作为量化器的输出，就实现了矢量量化。

在译码端有一个同样的码书，所以译码工作只是通过接收的码字序号在码书中搜索相应的码字，得到解码结果。

矢量量化的关键在于码书的构造，目前最流行的算法是由 Y. Linde、A. Buzo、R. M. Gray 共同提出的 LBG 算法，不需要知道输入矢量的概率分布，而是通过训练矢量集和采用

一定的迭代算法来逼近最优的再生码书。

① 采集用于构造码书的训练数据。为了得到性能较好的码书，一般要求训练样本的数量 N 和码字的个数 L 满足 $N \geqslant 20L$。

② 构造初始码书，常用的构造方法有随机码书法、白噪声码书法等。

③ 按照初始码书对所有的训练样本进行矢量量化，得到分属不同码字的 L 个样本集合和相应的量化误差。

④ 对每个样本集合进行聚类，根据聚类的结果修正初始码字，得到新的码书。

⑤ 判断量化误差是否小于门限值、迭代次数是否超出规定值，若是，则训练结束，否则转第③步继续。

矢量量化中第二个重要问题是量化误差的度量问题，即如何选择一个能准确反映量化前后失真大小的度量函数。这个问题与被编码对象的具体特性有关，没有一个统一的答案。

矢量量化的第三个问题是搜索运算量问题。在量化的过程中，需要比较待量化矢量与所有码字，计算相应的失真，然后选取失真最小的码字。当码书的规模较大时，需要很大的计算量。可以通过为码书安排某种特定的结构（如二叉树）来降低运算量。

矢量量化具有如下特点

① 样本数量 N 越大，量化失真越小。当 N 被设定后，如何找到误差最小的码书是矢量量化器设计的关键。

② 码字个数 L 越大，量化失真越小，但是相应的存储空间开销和搜索时间开销也越大。

③ 矢量维数 k 越大则编码效率越高，随之而来的是带来误差的危险也越大。

矢量量化算法简单，容易实现。矢量量化技术在图像压缩、语音编码、语音识别和数字水印等领域得到了广泛的应用。

6.5.2 预测编码

预测编码的概念最早是由 Peter Elias 在 1955 年提出的，是目前得到广泛应用的一种实用的熵压缩树码，是数据压缩理论的一个重要分支。预测编码根据离散信号之间在时间和空间上具有相关性的特点，解除信源的相关性，再进行编码。

1. 预测编码基本思想

预测编码的思路如下：利用前面的一个或多个信号对下一个信号进行预测，然后对实际值和预测值的差（预测误差）进行量化编码。编码器输出的不是样本的真实值，而是真实值与预测值的差。在译码端，译码器将接收到的这个差值与译码器的预测值相加，从而恢复信源符号。预测编码的编译码过程如图 6.4 和图 6.5 所示。

如果预测比较准确，那么误差信号就会很小。因此，在同等精度要求的条件下，可用较少的码符号表示差值，从而达到了压缩数据的目的。

根据预测器中预测值与信源符号过去时刻的值之间的函数关系，可将预测器分为线性预测器和非线性预测器，相应的预测编码称为线性预测编码和非线性预测编码。

设信源输出的符号序列为$\cdots, x_{-2}, x_{-1}, x_0, x_1, x_2, \cdots$，预测器根据所存储的过去的符号值对时刻 n 的符号进行预测，得到的预测值 x_n 与实际 y_n 之间的差值称为预测误差

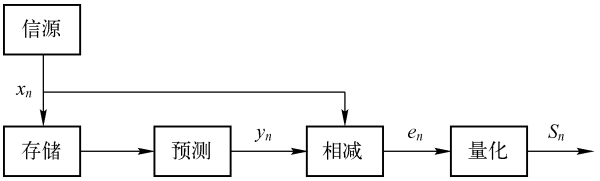


图 6.4　预测编码器

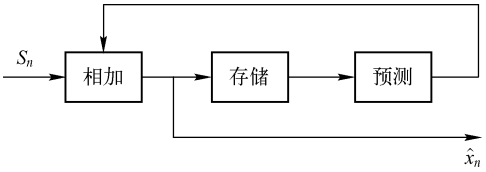


图 6.5　预测译码器

$$e_n = y_n - x_n$$

预测编码设计中的核心问题使如何选取预测函数以使预测误差满足某种最佳条件。常见的几种条件是最小均方误差、最小平均决定误差和最大零误差概率。不同准则下的最佳预测给出不同的预测值和不同的预测误差值，但是预测误差的分布与信源符号的一维分布具有相同的形状，其差别只是因为减去预测值所导致的分布的平移。

预测编码常用的是差分脉冲编码调制（Differential Pulse Code Modulation，DPCM）和自适应差分脉冲编码调制（Adaptive Differential Pulse Code Modulation，ADPCM）。由于声音和图像数据均由采样得到，且相邻值之间的差不会很大，可以用较少的位数来表示差值，因此适用于声音和图像数据的压缩。

2. 常用的编码方案

（1）差分脉冲调制编码（DPCM）

脉冲调制（Pulse Code Modulation，PCM）编码是将原始信号先经过时间采样，然后对每个样值进行量化后，作为数字信号的输出；而在差分脉冲调制编码（DPCM）中，为了压缩传输的数据，不是对每个样值都进行量化，而是预测下一样值，并量化实际值与预测值之间的差。解压缩时使用同样的预测器，将这个预测值与存储的已量化的差值相加，产生近似的原始信号，从而基本恢复原始数据。一般该编码每样值可压缩到 2 ~ 4 比特。

（2）自适应差分脉冲调制编码（ADPCM）

自适应差分脉冲调制编码（ADPCM）是根据信号分布不均匀的特点，随输入信号的变化而自动改变量化区间的大小，并选择预测参数。系统对输入信号的变化采用不同的量化区间（自适应量化）：对于能量分布较大的系数分配较多的比特数，采用较小的量化步长；反之分配较小的比特数，采用较大的量化步长，从而达到压缩的目的。

（3）帧间预测编码

对于动态图像，帧间预测编码利用帧间的时间相关性进一步消除图像的冗余度，基于预测技术来提高动态图像的压缩比。

6.5.3 变换编码

1. 变换编码基本思想

预测编码主要是在时域消除信源的相关性，而变换编码则主要是在变换域上消除信源的相关性。变换编码常用的变换有傅里叶变换、余弦变换、离散余弦变换（Discrete Cosine Transform，DCT）、K－L 变换、沃尔什变换、小波变换等。

变换编码是先对信号进行变换，从一种信号空间变换为另一种信号空间，然后针对变换后的信号进行编码。一般是把分布在时空域的信号（如时域的语音信号和平面空间的图像信号）映射到变换域（如频域的频谱信号和其它正交矢量空间域），原来相关性很强的原始信号经过变换后，得到的变换域系数相互独立，并且能量集中在少数几个变换系数上。这样，只需对这少量的系数进行量化编码，达到数据压缩的目的。

变换编码最典型的例子就是傅里叶变换，它将时域信号变换成频谱信号。将时域信号变换到频域信号，只需振幅和频率两个参数，数据相关性减少（样值更具独立），而且声音、图像的大部分信号都是低频信号，在频域中信号的能量较集中，因此再对这些变换参数进行采样、量化、编码处理，即可压缩数据。

通常，变换编码的基本过程如下。

① 变换：将时域信号映射到变换域。

② 变换域采样：对变换系数进行采样。变换后的样值具有有序性和独立性。

③ 量化编码：要求使数据量尽可能减少、量化失真也最小。

2. 常用的编码方案

常用的编码方案中，K－L 变换编码去相关性最好，但算法复杂、实现困难；离散余弦变换的性能接近 K－L 变换，也容易实现，被选为众多图像压缩编码技术标准的基本算法。

（1）K－L 变换编码

K－L 变换又称为霍特林（Hotelling）变换，是均方差意义下的最佳变换，是在已知输入信号矩阵的条件下，根据其协方差矩阵去寻找另一种正交变换，使变换后的协方差矩阵成为或接近为一个对角矩阵。

K－L 变换以原始数据的协方差矩阵的归一化正交特征向量构成的正交矩阵作为变换矩阵，对原始数据进行正交变换，在变换域上实现数据压缩。

假设向量集合$\{\boldsymbol{x}_i\}(i=1,2,\cdots)$中的$\boldsymbol{x}$可以用变换矩阵的基向量$\boldsymbol{u}_j(j=1,2,\cdots)$来展开，即

$$\boldsymbol{x}=\sum_{j=1}^{\infty}c_j\boldsymbol{u}_j$$

基向量满足正交性

$$\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_j=\begin{cases}1 & j=i\\ 0 & j\neq i\end{cases}$$

在离散情况下，$\boldsymbol{x}$可使用有限基向量集合来近似，即

$$\hat{\boldsymbol{x}} = \sum_{j=1}^{d} c_j \boldsymbol{u}_j$$

其均方误差为

$$\xi = E[(\boldsymbol{x} - \hat{\boldsymbol{x}})^{\mathrm{T}}(\boldsymbol{x} - \hat{\boldsymbol{x}})]$$

$$= E\left[\left(\sum_{j=d+1}^{\infty} c_j \boldsymbol{u}_j\right)^{\mathrm{T}}\left(\sum_{j=d+1}^{\infty} c_j \boldsymbol{u}_j\right)\right]$$

$$= E\left[\sum_{j=d+1}^{\infty} c_j^2\right]$$

将展开式系数（可理解为 $\boldsymbol{x}$ 在基坐标上的投影，而展开式系数就是坐标值）$c_j = \boldsymbol{u}_j^{\mathrm{T}}\boldsymbol{x}$ 代入均方误差表达式，则

$$\xi = E\left[\sum_{j=d+1}^{\infty} \boldsymbol{u}_j^{\mathrm{T}}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{u}_j\right]$$

$$= \sum_{j=d+1}^{\infty} \boldsymbol{u}_j^{\mathrm{T}} E(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})\boldsymbol{u}_j$$

$$= \sum_{j=d+1}^{\infty} \boldsymbol{u}_j^{\mathrm{T}}\boldsymbol{\psi}\boldsymbol{u}_j$$

式中，$\boldsymbol{\psi} = E(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})$ 为样本的自相关矩阵。用拉格朗日条件极值法求均方误差的极小值，可得 $(\boldsymbol{\psi} - \lambda_j E)\boldsymbol{u}_j = 0 (j = d+1, d+2, \cdots)$，其解就是使均方误差为极小的基向量 $\boldsymbol{u}_j$。$\boldsymbol{u}_j$ 为矩阵 $\boldsymbol{\psi}$ 的特征向量，其对应的特征值为 λ_j，则截断均方误差为

$$\xi = \sum_{j=d+1}^{\infty} \lambda_j$$

式中，λ_j 为矩阵 $\boldsymbol{\psi}$ 的特征值。

如果能够取矩阵 $\boldsymbol{\psi}$ 的 d 个最大特征值 $\lambda_j (j = 1, 2, \cdots, d)$，所对应的特征向量 $\boldsymbol{u}_j (j = 1, 2, \cdots, d)$ 来构成 K－L 变换坐标系，便可以实现 D 维空间到 $d (d \leqslant D)$ 维空间的压缩映射，并且均方误差最小。K－L 变换可以用较少数量的特征对样本进行描述，在人脸识别、图像压缩和信号传输等多种应用中用于特征提取和降维。

K－L 变换虽然是均方误差意义下的最佳变换，是最能去除原始数据之间相关性的一种变换，但需要先知道信源的协方差矩阵并求出特征值。求特征值与特征向量并不是一件容易的事，维数较高时甚至求不出来。即使能借助计算机求解，也很难满足实时处理的要求，从编码应用看，还需要将这些信息传输给接收端。这是 K－L 变换在工程实践中不能广泛使用的原因。人们一方面继续寻求解特征值与特征向量的快速算法，另一方面寻找一些虽不是“最佳”也有较好性能且容易实现的一些变换方法，K－L 变换就常常作为这些变换性能的评价标准。

（2）离散余弦变换编码

离散余弦变换（DCT）编码是从快速傅里叶变换（Fast Fourier Transform，FFT）中取实部，再利用余弦正交变换算法对不同的信号进行压缩，广泛地应用在图像压缩中。

离散傅里叶变换（Discrete Fourier Transform，DFT）尽管通过快速傅里叶变换（FFT）可以

提高运算速度，但是需要进行复数运算，在图像编码特别是在实时处理中非常不便。离散傅里叶变换在实际的图像通信系统中很少使用，但具有理论的指导意义。根据离散傅里叶变换的性质，实偶函数的傅里叶变换只包含实的余弦项，因此构造了一种实数域的变换——离散余弦变换(DCT)。通过研究发现，离散余弦变换除了具有一般的正交变换性质外，其变换阵的基向量很近似于 Toeplitz 矩阵的特征向量，后者体现了人类的语言、图像信号的相关特性。因此，在对语音、图像信号变换的确定的变换矩阵正交变换中，离散余弦变换变换被认为是一种准最佳变换。在近年颁布的一系列视频压缩编码的国际标准建议中都把离散余弦变换作为其中的一个基本处理模块。

离散余弦变换除了实数变换、确定的变换矩阵、准最佳变换性能等特点外，二维离散余弦变换还是一种可分离的变换，可以用两次一维变换得到二维变换结果。

离散余弦变换有 8 种标准类型，其中 4 种是常见的。最常用的是离散余弦变换，通常我们所说的离散余弦变换指的是下面这种。

正变换：

$$y(k) = a_k \sqrt{\frac{2}{N}} \sum_{n=0}^{N-1} x(n) \cos \frac{(2n+1)k\pi}{2N} \qquad k = 0,1,2,\cdots,N-1$$

逆变换：

$$x(n) = \sqrt{\frac{2}{N}} \sum_{k=0}^{N-1} a_k y(k) \cos \frac{(2n+1)k\pi}{2N} \qquad k = 0,1,2,\cdots,N-1$$

式中，

$$a_k = \begin{cases} \dfrac{1}{\sqrt{2}} & k = 0 \\ 1 & k = 1,2,\cdots,N-1 \end{cases}$$

离散余弦变换经常被信号处理和图像处理使用，用于对信号和图像（包括静止图像和运动图像）进行有损数据压缩。这是由于离散余弦变换具有很强的能量集中特性：大多数的自然信号（包括声音和图像）的能量都集中在离散余弦变换后的低频部分，而且当信号具有接近马尔科夫过程的统计特性时，离散余弦变换的去相关性接近于 K－L 变换的性能。

动手实践：图像的离散余弦变换

利用 MATLAB 工具箱中的离散余弦变换函数，对一个图像实现离散余弦变换，并求出逆变换后重构图像的均方误差。

输入：读入一幅图像。

输出：离散余弦变换结果、逆变换后重构图像和均方误差。

习　题　6

6.1　根据率失真函数的性质，画出一般 $R(D)$ 曲线并说明其物理意义。为什么它是非负且非增的？

6.2　设输入符号集为 $X = \{0,1\}$，输出符号集为 $Y = \{0,1\}$。定义失真函数为

$$d(0,0)=d(1,1)=0$$
$$d(0,1)=d(1,0)=1$$

试求失真矩阵 $\boldsymbol{D}$。

6.3 设输入符号集与输出符号集为 $X=Y=\{0,1,2,3\}$，输入信源的分布为 $P(X=i)=\frac{1}{4}(i=1,2,3,4)$。设失真矩阵为

$$\boldsymbol{D}=\begin{bmatrix}0&1&1&1\\1&0&1&1\\1&1&0&1\\1&1&1&0\end{bmatrix}$$

求 $D_{\min}$ 和 $D_{\max}$ 及 $R(D)$。

6.4 设二进制信源为 $\begin{bmatrix}X\\P\end{bmatrix}=\begin{bmatrix}0&1\\\frac{1}{2}&\frac{1}{2}\end{bmatrix}$，失真函数矩阵为 $\boldsymbol{D}=\begin{bmatrix}0&a\\a&0\end{bmatrix}$。求这个信源的 $D_{\min}$ 和 $D_{\max}$ 及率失真函数 $R(D)$。

6.5 设二元等概离散无记忆信源，通过一个二进制对称信道 $\boldsymbol{P}=\begin{bmatrix}1-\varepsilon&\varepsilon\\\varepsilon&1-\varepsilon\end{bmatrix}$。其失真函数 d_{ij} 定义为

$$d(i,j)=\begin{cases}0 & i=j\\1 & i\neq j\end{cases}$$

（1）求失真矩阵 $\boldsymbol{D}$。

（2）求平均失真度 $\overline{D}$。

（3）求率失真函数 $R(D)$ 的定义域。

6.6 证明对离散信源，$R(D=0)=H(X)$ 的充要条件是失真矩阵的每行至少有一个 0，而每列至多有一个 0。

6.7 失真函数矩阵为 $\boldsymbol{D}=\begin{bmatrix}1&2\\3&2\end{bmatrix}$，对于一个等概分布的伯努利随机变量，求 $R(D)\in(0,1)$ 对应的定义域。

6.8 三元信源的概率分布为 $\{0.4,0.4,0.2\}$，失真函数为

$$d(i,j)=\begin{cases}0 & i=j\\1 & i\neq j\end{cases}$$

（1）求率失真函数 $R(D)$。

（2）计算此信源分别用容量为 1 比特/符号和 0.1 比特/符号的信道传输时的平均失真。

6.9 设已知离散无记忆信源在给定失真量度 $d(i,j)(i=1,2,\cdots,r;\ j=1,2,\cdots,s)$ 下的信息率失真函数为 $R(D)$，现定义新的失真量度 $d'(i,j)=d(i,j)-g_i$。试证明：在新的失真量度下信息率失真函数 $R'(D)$ 为 $R'(D)=R(D+G)$，其中 $G=\sum_i p(a_i)g_i$。

6.10 设无记忆信源$\begin{bmatrix} X \\ p(x) \end{bmatrix}=\begin{bmatrix} -1 & 0 & 1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$，接收符号 $A_Y=\left\{-\frac{1}{2},\frac{1}{2}\right\}$，失真矩阵 $\boldsymbol{D}=\begin{bmatrix} 1 & 2 \\ 1 & 1 \\ 2 & 1 \end{bmatrix}$，试求 $D_{\max}$ 和 $D_{\min}$ 及达到 $D_{\max}$、$D_{\min}$ 时的转移概率矩阵。

6.11 失真测度 $d(x,\hat{x})$ 定义为

$$\boldsymbol{D}=\begin{bmatrix} 0 & 1 & \infty \\ \infty & 1 & 0 \end{bmatrix} \qquad x=0,1;\hat{x}=0,\varepsilon,1$$

设随机变量 X 的概率分布为$\begin{bmatrix} X \\ P \end{bmatrix}=\begin{bmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$。求：

（1）$D_{\max}$ 和 $D_{\min}$ 的值。

（2）$R(D_{\min})$ 和 $R(D_{\max})$ 的值。

6.12 某三元信源$\begin{bmatrix} X \\ P \end{bmatrix}=\begin{bmatrix} 0 & 1 & 2 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$，失真矩阵为 $\boldsymbol{D}=\begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \\ 3 & 2 & 1 \end{bmatrix}$。求：

（1）$D_{\max}$ 和 $D_{\min}$ 的值。

（2）达到 $D_{\max}$ 和 $D_{\min}$ 的信道转移概率矩阵 $\boldsymbol{P}$。

（3）$R(D_{\min})$ 和 $R(D_{\max})$ 的值。

第 7 章　信息论的应用

信息论有狭义和广义之分。狭义信息论即香农早期的研究成果，以编码理论为中心，主要研究通信系统模型、信息的度量、信息容量、编码理论及噪声理论等。广义信息论又称为信息科学，主要研究信息处理过程中的基本理论，包括文字处理、图像识别、学习理论及其各种应用，是狭义信息论在各领域的应用和推广。20 世纪 70 年代以来，电视、数据通信、遥感和生物医学工程的发展，向信息科学提出大量的研究课题，如信号的压缩、增强、恢复等图像处理和传输技术，信号的特征抽取、分类和模式识别的理论和方法，出现了实用的图像处理和模式识别系统。本章介绍几种在信息采集和处理过程中的信息论应用。

7.1　最大熵谱估计

随机信号的持续时间无限长，是具有无限大能量的功率信号，不满足傅里叶变换条件，而且不存在解析表达式，因此不能够用确定信号的频谱计算方法去分析它的频谱。然而，虽然它的频谱不存在，但其自相关函数是确定函数。对于平稳随机信号，它的自相关函数的傅里叶变换就是它的功率谱密度函数，简称功率谱。因此，通过分析功率谱密度可以了解这个随机信号的频谱分布。通常是利用给定的 N 个样本数据估计这个平稳随机信号的功率谱密度，叫做功率谱估计(PSD)。功率谱估计可以分为经典功率谱估计和现代功率谱估计。

经典谱估计（线性、非参数化方法）包括周期图法、自相关（Bartlett，BT）法以及它们的修正方法。经典谱估计以傅里叶变换为基础，计算效率高，但是因为观察数据只有有限个，观察不到的数据被认为是 0，以及给数据加窗，因此存在频率分辨率低、旁瓣泄漏、谱估计方差性能不好等严重的缺陷。为此，人们在提高功率谱估计的分辨率方面提出了很多新方法。

现代谱估计（非线性、参数化方法）是通过观测数据来估计信号模型的参数，再通过求信号模型的输出功率估计原信号的功率谱，对短序列的估计精度高。常用模型有 ARMA 模型、AR 模型、MA 模型。由于 AR 模型具有良好的性能，因此被研究最多也得到最广泛的应用。此外，现代谱估计还包括最大似然谱估计、最大熵谱估、特征分解法谱估计等。

最大熵谱估计法（Maximun Entropy Spectral Estimation，MESE）是 1967 年由 J. P. Burg 提出的，是将已知的有限长度的自相关序列以外的数据用外推法求得（因为观察不到的数据和观察数据具有相关性），不是把它们当作零。假设已知 N 个自相关函数的值$R_{xx}(0)$，$R_{xx}(1),\cdots,R_{xx}(N)$，按什么原则外推 $R_{xx}(N+1)$，$R_{xx}(N+2),\cdots$呢？在保证自相关函数的 Toeplitz 矩阵是正定的情况下有无穷多种外推法，Burg 认为，外推的自相关函数应使时间序列表现出最大熵，因此这种方法被称为最大熵谱估计法。

熵是不确定性的度量，最大熵为最大不确定度，即它的时间序列最具随机性，它的 PSD

应是最平滑（最白色）。对于离散随机序列，如果随机变量取连续值，它的概率密度函数用联合概率密度函数代替。

对于 N 维高斯分布，它的自相关函数矩阵为

$$\boldsymbol{R}_{xx}(N)=\begin{bmatrix} R_{xx}(0) & R_{xx}(-1) & \cdots & R_{xx}(-N) \\ R_{xx}(1) & R_{xx}(0) & \cdots & R_{xx}(-(N-1)) \\ \vdots & \vdots & \ddots & \vdots \\ R_{xx}(N) & R_{xx}(N-1) & \cdots & R_{xx}(0) \end{bmatrix}$$

它的熵为

$$H=\ln\left[(2\pi e)^{\frac{N}{2}}(\det\boldsymbol{R}_{xx}(N))^{\frac{1}{2}}\right]$$

其中，$\det\boldsymbol{R}_{xx}(N)$代表自相关函数矩阵的行列式。要使熵 H 最大，就要求 $\det\boldsymbol{R}_{xx}(N)$最大。

现已知 $R_{xx}(0),R_{xx}(1),\cdots,R_{xx}(N)$，求 $R_{xx}(N+1)$，并且最大熵原则要求 $R_{xx}(N+1)$须使得 $\det\boldsymbol{R}_{xx}(N+1)$最大。

由于自相关函数的矩阵必是正定的，故 $\det\boldsymbol{R}_{xx}(N+1)$必大于零，即

$$\det\boldsymbol{R}_{xx}(N+1)=\begin{bmatrix} R_{xx}(0) & R_{xx}(1) & \cdots & R_{xx}(N+1) \\ R_{xx}(1) & R_{xx}(0) & \cdots & R_{xx}(N) \\ \vdots & \vdots & \ddots & \vdots \\ R_{xx}(N+1) & R_{xx}(N) & \cdots & R_{xx}(0) \end{bmatrix}>0$$

解方程

$$\frac{d}{\mathrm{d}R_{xx}(N+1)}\det\boldsymbol{R}_{xx}(N+1)=0$$

可求得使 $\det\boldsymbol{R}_{xx}(N+1)$最大的 $R_{xx}(N+1)$，它满足下列方程：

$$\begin{bmatrix} R_{xx}(1) & R_{xx}(0) & \cdots & R_{xx}(N-1) \\ R_{xx}(2) & R_{xx}(1) & \cdots & R_{xx}(N-2) \\ \vdots & \vdots & \ddots & \vdots \\ R_{xx}(N+1) & R_{xx}(N) & \cdots & R_{xx}(1) \end{bmatrix}=0$$

上式是 $R_{xx}(N+1)$的一次函数，由此式可解出 $R_{xx}(N+1)$。

用类似方法求得 $R_{xx}(N+2)$，以此类推。这样每步都按最大熵的原则外推后一个自相关序列的值，可以外推到任意多个而不必认为它们是零。这就是最大熵谱估计法的基本思想。

将信息论中最大熵原理应用于时间序列的功率谱估计，是频谱分析上的重要进展，因其分辨率高，在众多领域已经得到广泛应用。例如，在分析地震前兆信号的频谱变化中，由于震源体在孕震过程中可能出现多种地球物理和地球化学变化，用高分辨率的最大熵谱估计方法，可以对监测得到的时间序列和空间序列进行分析，得到前兆信号的特征频率成分。在地球物理数据处理中，由于许多地球物理场（如地磁场、地电场、重力场、固体潮汐与地壳形变以及地球的自由振荡等）呈现出明显的时空周期特征，最大熵谱估计分辨力高，可以对这些周期特征进行精确分析。另外，最大熵谱估计还广泛用于降噪，以提高信噪比。

7.2 基于信息论的信息融合技术

信息融合又称为数据融合，也可以称为多传感器信息融合。信息融合充分利用多个传感

器资源，通过对这些传感器及其观测信息的合理支配和使用，把多个传感器在时间或空间上的冗余或互补信息依据某种准则来进行组合，以获得被测对象的一致性解释或描述，使该信息系统由此而获得比它的各组成部分的子集所构成的系统更优越的性能。

按照数据抽象的不同层次，融合可分为三层，即数据层融合、特征层融合和决策层融合，如图 7.1 所示。

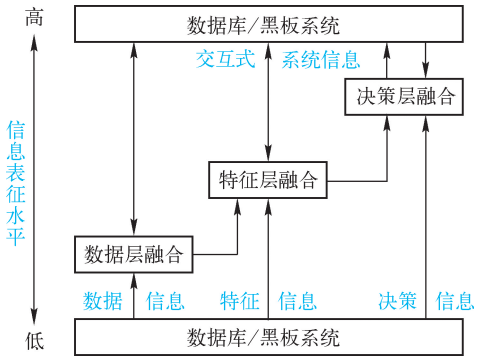


图 7.1　多传感器融合层次化结构

数据层融合是指在原始数据层上进行的融合，即各种传感器对原始信息未作很多预处理之前就进行的信息综合分析，这是最低层次的融合。

特征层合属于中间层次，对来自传感器的原始信息进行特征提取，然后对特征信息进行综合分析和处理。特征层融合可划分为两类：目标状态信息融合和目标特性融合。目标状态信息融合主要用于多传感器目标跟踪领域。融合系统首先对传感器数据进行预处理以完成数据校准，然后主要实现参数相关和状态向量估计。目标特性融合就是特征层联合识别，具体的融合方法仍是模式识别的相应技术，只是在融合前必须先对目标特征进行相关处理，把特征向量分类成有意义的组合。

决策层合是一种高层次融合，其结果为指挥控制决策提供依据。因此，决策级融合必须从具体决策问题的需求出发，充分利用特征融合所提取的测量对象的各类特征信息，采用适当的融合技术来实现。决策级融合是三级融合的最终结果，直接针对具体决策目标，融合结果直接影响决策水平。但是，决策级融合首先要对原传感器信息进行预处理以获得各自的判定结果，所以预处理代价高。

信息融合首先在军事领域取得了成功的应用，目前在许多民用领域也取得了很大的进展，包括机器人和智能仪器系统、智能制造系统、战场任务与无人驾驶飞机、航天应用、目标检测与跟踪、图像分析与理解、惯性导航、模式识别等领域。

多传感器信息融合的常用算法总的来说分为基于物理模型、基于参数分类、基于认知模型三类。基于参数分类算法又分为统计法和信息论方法，其中信息论方法包括聚类、神经网络和熵法等。

7.2.1　聚类分析法

聚类分析型信息融合技术是以统计聚类分析或模糊聚类分析原理为基础，在多目标、多

传感器、大量观测数据样本的情况下，使来自同一目标的数据样本自然聚集、来自不同目标的数据样本自然隔离，从而实现多目标信息融合。聚类分析法是一类方法的总称，包括5种。

1. 基于层次的聚类（hierarchical methods）

基于层次的聚类有自下而上法（bottom－up）和自上而下法（top－down）两种路径。自下而上法就是一开始每个个体（object）都是一个“类”，然后寻找同类，最后形成几个“类”。自上而下法则相反，一开始所有个体都属于一“类”，然后排除异己，最后形成几个“类”。在实际应用时要根据数据特点以及想要的“类”的个数，来考虑是自上而下更快还是自下而上更快。判断是否同类的方法包括最短距离法、最长距离法、中间距离法、类平均法等。

2. 基于划分的聚类（partition－based methods）

简单来说，想象有一堆散点需要聚类，想要的聚类效果就是“类内的点都足够近，类间的点都足够远”。首先，要确定这堆散点最后聚成几类，然后挑选几个点作为初始中心点，再依据预先定好的启发式算法（heuristic algorithms）给数据点进行迭代重置（iterative relocation），直到最后到达“类内的点都足够近，类间的点都足够远”的目标效果。正是根据“启发式算法”，形成了k－means算法及其变体包括k－medoids、k－modes、k－medians、kernel k－means等算法。

3. 基于密度的聚类（density－based methods）

基于密度的聚类针对不规则形状的聚类。简单来说，其原理就是画圈儿，要定义两个参数：圈的最大半径，圈中最少应容纳的点数。最后在一个圈里的就是一个类。DBSCAN（Density－Based Spatial Clustering of Applications with Noise）是其中的典型，对噪声数据的处理也比较好，但是对这两个参数的设置是个问题，对参数的设置非常敏感。

4. 基于网格的聚类（grid－based methods）

基于网格的聚类的原理就是，将数据空间划分为网格单元，将数据对象集映射到网格单元中，并计算每个单元的密度。根据预设的阈值判断每个网格单元是否为高密度单元，由邻近的稠密单元组形成“类”。该方法的优点就是执行效率高，因为其速度与数据对象的个数无关，而只依赖于数据空间中每维单元的个数。但缺点也不少，如对参数敏感、无法处理不规则分布的数据、维数灾难等。STING（STatistical INformation Grid）和CLIQUE（CLustering In QUEst）是该方法的代表性算法。

5. 基于模型的聚类（model－based methods）

基于模型的聚类主要是指基于概率模型的方法和基于神经网络模型的方法，尤其以基于概率模型的方法居多。这里的概率模型主要指概率生成模型，同一“类”的数据属于同一种概率分布。该方法的优点就是对“类”的划分不那么确定，而是以概率形式表现，每类的特征也可以用参数来表达；但缺点就是执行效率不高，特别是分布数量很多并且数据量很

少的时候。其中，最典型、最常用的方法就是高斯混合模型（Gaussian Mixture Models，GMM）。基于神经网络模型的方法主要就是指 SOM（Self Organized Maps）。

7.2.2 神经网络法

神经网络也称为人工神经网络（Artificial Neural Network，ANN）是基于现代神经生物学和认知科学在信息处理领域应用的研究成果，通过大量功能简单的神经元，完成复杂的信息处理。

人工神经网络具有很强的容错性以及自学习、自组织及自适应能力，能够模拟复杂的非线性函数关系。神经网络的这些特性以及它具有的大规模并行模拟处理能力，恰好满足了多传感器信息融合技术处理的要求。人工神经网络可以避开模式识别方法中建模和特征提取的过程，并能实现实时识别，因此用于信息融合可以得到较为理想的结果。

神经网络就像一个刚开始学习认知的小孩子开始认东西，第一天，他看见一只哈巴狗，大人告诉他这是狗；第二天他看见一只猫，他开心地说，这是狗，大人纠正他，这是猫；第三天，他看见一只京巴狗，他有点迷惑了，大人告诉他这是狗……直到有一天，他可以分清任何一只猫或者狗。

神经网络就是在模拟人的大脑，把每个节点当成一个神经元，这些“神经元”组成的网络就是神经网络。而由于计算机出色的计算能力和细节把握能力，在大量的训练数据的基础上，神经网络往往有比人更出色的表现。

当然，也可以把神经网络当成一个黑箱子，只要告诉它输入、输出，它可以学到输入与输出的函数关系，可以逼近任意的函数。所以在理论上，只要数据量够大，“箱子容量”够大（神经元数量），神经网络就可以学到你要的东西。

人工神经网络算法中应用最为广泛的是 1986 年由 Rumelhart 和 McCelland 为首的科学家小组提出的 BP（Back Propagation，反向传播）神经网络。BP 神经网络通常由输入层（Input Layer）、隐层（Hidden Layer）以及输出层（Output Layer）构成。其中，输入层和输出层的节点数由训练样例的输入和输出决定，隐层可以由多层构成，隐层节点数则由训练者通过某种方法决定。最简单的是三层 BP 神经网络如图 7.2 所示。

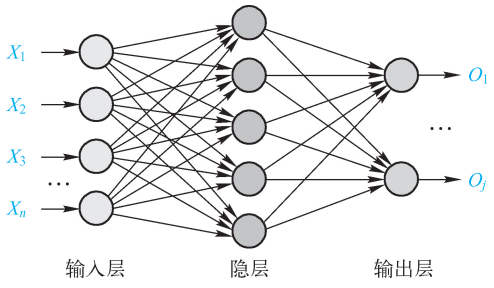


图 7.2 三层神经网络

BP 神经网络的训练思路是前馈训练——误差反向传播算法，学习过程由信号的正向传播与误差的反向传播两个过程组成。正向传播时，输入样本从输入层传入，经各隐层逐层处理后，传向输出层。若输出层的实际输出与期望的输出（教师信号）不符，则转入误差的

反向传播阶段。反向传播时，将输出以某种形式通过隐层向输入层逐层反传，并将误差分摊给各层的所有单元，从而获得各层单元的误差信号，此误差信号即作为修正各单元权值的依据。若训练结果与实际的输出存在误差，则根据误差梯度下降的原则通过不同的隐层，并修正各层的权值。通过循环的正向训练和反向传播的过程，逐步修正各隐层的权值，从而最终达到拟合训练目标的目的。

BP 神经网络的思想可以总结为利用输出后的误差来估计输出层的直接前导层的误差，再用这个误差估计更前一层的误差，如此一层一层地反传下去，就获得了所有其他各层的误差估计。BP 神经网络利用一种称为激活函数来描述层与层输出之间的关系，从而模拟各层神经元之间的交互反应。激活函数必须满足处处可导的条件，比较常用的是一种称为 S 形函数的激活函数

$$f(x)=\frac{1}{1+e^{-x}}$$

通常 BP 神经网络的训练结束条件有以下几类：① 达到最大的迭代次数；② 神经网络的输出误差达到设计需求（通常误差采用最小均方误差 MSE 确定）；③ 连续迭代规定代数但训练误差不发生显著改变。

BP 神经网络可以以任意精度逼近任何非线性函数而无需事先描述该函数，具有一定的抗噪能力或训练实例的容错性，能够在一定程度上对未分类实例进行区分。其缺陷主要有：容易陷入局部极小值，BP 神经网络的结构不唯一，算法收敛较慢等。

7.2.3 熵法

信息融合的本质是一个由底至顶对多源信息逐层进行整合的信息处理过程。对多源信息的逐层整合意味着在更高层次上信息的不确定性减少，描述不确定的强有力工具就是香农的信息熵。融合是对同一表达层次上多源信息的合成，这样就可以把输入信息和输出信息用两个概率空间及其上定义的信息熵来描述，由此分析融合过程中信息的传递和转换，揭示融合的本质，指导融合系统设计，并且作为融合效果的评估指标。

熵法从信息论的观点出发，导出多传感器信息融合系统中信息的冗余性与互补性的定量描述，并从该角度出发，利用最小条件熵等关于信息熵的准则来解决多传感器信息融合中的目标识别问题，该方法的主要优点是可以充分利用多传感器信息，使融合系统获得的信息量最大。

熵法在概念上也许是最简单的方法，但是由于要对传感器输入加权以及应用了阈值和其他判定逻辑，从而增加了算法的复杂性。尽管如此，熵法仍具有应用价值，尤其对于实时性要求很强的系统，当准确的先验统计不可利用时，或者从整个成本效益观点来看，其具有很大的应用空间。

7.3 压缩感知与信息论

压缩感知（Compressed sensing）也被称为压缩采样（Compressive sampling）、稀疏采样（Sparse sampling）、压缩传感，是 E. J. Candes、J. Romberg、T. Tao 和 D. L. Donoho 等科学家

于 2004 年提出的。作为一个新的采样理论，压缩感知通过开发信号的稀疏特性，在远小于奈奎斯特采样率的条件下，用随机采样获取信号的离散样本，然后通过非线性重建算法完美的重建信号。压缩感知理论一经提出，就在图像处理、地球物理、医学成像、计算机科学、信号处理、应用数学等领域受到高度关注。

信号采样是模拟的物理世界通向数字的信息世界的必由之路。奈奎斯特采样定理告诉我们，只有当采样速率达到信号带宽的 2 倍以上时，才能由采样信号精确重建原始信号，带宽是奈奎斯特采样定理对采样的本质要求。但是随着信号的带宽变得越来越大，人们对信号的采样速率、传输速度和存储空间的要求也变得越来越高。为了缓解对信号传输速度和存储空间的压力，当前常见的解决方案是信号压缩，如基于小波变换的 JPEG2000 标准。但是，信号压缩实际上是一种严重的资源浪费，因为大量的采样数据在压缩过程中被丢弃了，而它们对于信号来说是不重要的或者只是冗余信息。从这个意义而言，我们得到以下结论：带宽不能本质地表达信号的信息，基于信号带宽的奈奎斯特采样机制是冗余的或者说是非必要的。

压缩感知理论指出：与信号带宽相比，稀疏性能够直观地而且相对本质地表达信号的信息。当信号在某个变换域是稀疏的或可压缩的，可以利用与变换矩阵不相关的测量矩阵将变换系数线性投影为低维观测向量，同时这种投影保持了重建信号所需的信息，通过进一步求解稀疏最优化问题，就能够从低维观测向量精确地或高概率精确地重建原始高维信号。在该理论框架下，采样速率不再取决于信号的带宽，而在很大程度上取决于两个基本准则：稀疏性和不相关性，或者稀疏性和等距约束性。因此，压缩感知理论是一种基于信息的采样理论框架，使得采样过程既能保持信号信息，又能用远少于奈奎斯特采样定理所要求的采样数目就可精确或近似精确重建原始信号，而且同时实现信号的采样与压缩，抛弃了当前信号采样中的冗余信息，直接从连续时间信号变换得到压缩样本

压缩感知理论的核心思想主要包括三点。

① 信号本身应具有稀疏性。传统的香农信号表示方法只开发利用了最少的被采样信号的先验信息，即信号的带宽。但是现实生活中，很多信号具有稀疏的结构特点，相对于带宽信息的自由度，这些结构特点使得信号可以由更少的自由度所决定。换句话说，在很少的信息损失情况下，这种信号可以用很少的数字编码表示。所以，在这种意义上将，这种信号是稀疏信号（或者近似稀疏信号、可压缩信号）。

② 非相关观测。理论证明压缩感知的采样方法只是一个简单地将信号与一组确定的波形进行相关的操作。这组波形要求是与信号所在的稀疏空间不相关的。通过这样一个非自适应的采样方法获取到稀疏信号的有用信息，将信号压缩成较小的样本数据。

③ 需要一个快速鲁棒的信号重建算法将稀疏信号还原为原始信号。非线性优化重构算法利用少量的压缩采样值中重构原始信号，在尽量保证不失真的前提下，将低速采样得到的低维信号恢复成高维信号，是压缩感知的核心问题。E. J. Candes 和 T. Tao 等人的工作指出，如果假定信号（无论是图像还是声音还是其他种类的信号）满足某种特定的稀疏性，那么从这些少量的测量数据中，确实有可能还原出原始的较大的信号来，其中所需的计算部分是一个复杂的迭代优化过程，即所谓 L1 – 最小化算法。目前，重构算法主要有 3 种：贪婪算法、组合算法和凸松弛法。

在信息论中，压缩的前提是数据的信息之中存在着冗余。所谓信息冗余，就是可以确定或者可以根据其他信息推测出的数据，如果能将这种数据全部去除，只保留无法根据其他信

息确定的信息，那么就实现了数据的压缩。数据压缩的过程就是让每个符号尽可能等概分布。

数据压缩在压缩感知中的实现，就是在可接受的信息熵损失范围内，对信号进行某个变换域处理，并且变换之后的信号是稀疏的。所以压缩感知第一步需要做的，就是找到这样一个稀疏域，而找到稀疏域过程中最为关键的一点是，找到或者构建适合该信号的正交基底来表示该信号。对于不同类型的原始信号来说，就是找出一本能够根据信号类型选择合适正交基底的字典。

压缩感知使得我们可以在采集数据的时候只简单采集一部分数据，然后把复杂的部分交给数据还原的这一端来做，这正好符合常规采集设备和处理（解压缩）设备的性能要求。通常情况下，采集设备往往是廉价、省电、计算能力较低的便携设备，如傻瓜相机或者录音笔等，而处理（即解压缩）信息的过程往往在计算机上进行，它有更高的计算能力，也常常没有便携和省电的要求。而传统的数据采集 – 压缩 – 传输 – 解压缩的模式却是用廉价节能的设备来处理复杂的计算任务，而用大型高效的设备处理相对简单的计算任务，以音频压缩为例，压制一个 MP3 文件的计算量远大于播放（即解压缩）一个 MP3 文件的计算量。

在医学图像领域里，这个方案特别有好处，因为采集数据的过程往往是对病人带来很大麻烦甚至身体伤害的过程。以 X 光断层扫描为例，X 光辐射会对病人造成身体损害，而压缩感知意味着我们可以用比经典方法少得多的辐射剂量来进行数据采集，这在医学上的意义是不言而喻的。

这个思路可以扩展到很多领域。在大量的实际问题中，我们倾向于尽量少地采集数据，或者由于客观条件所限不得不采集不完整的数据。如果这些数据和我们所希望重建的信息之间有某种全局性的变换关系，并且我们预先知道那些信息满足某种稀疏性条件，就总可以试着用类似的方式从比较少的数据中还原出比较多的信号来。

总之，压缩感知基于信号的可压缩性，从尽量少的数据中提取尽量多的信息，毫无疑问是一种有着极大理论和应用前景的想法。可以这么说，信息论的一些基本概念和原理（如信源、信道、信源编码、信道编码、率失真、费诺不等式、数据处理定理等）为压缩感知研究尤其是在性能限（如采样数）的界定等方面提供了理论基础；另一方面，压缩感知提供了采集、存储、传输、恢复稀疏信号的高效方法，以其独特的理念和算法模式，提供了直接对信息进行采样和处理的机制，延拓了经典信息论的范畴。压缩感知是传统信息论的一个延伸，但是又超越了传统的压缩理论，成为了一个崭新的分支。压缩感知从诞生之日起到现在不过十余年时间，其影响却已席卷了大半个应用科学。

附录 A　信息论学习要点

1. 自信息

自信息表示随机事件 x_i 发生的不确定性或发生所含有的信息量。自信息量 $I(x_i)$ 定义为

$$I(x_i) = -\log p(x_i) = \log \frac{1}{p(x_i)}$$

式中，$p(x_i)$ 为该事件发生的概率。

2. 互信息

互信息 $I(x_i;y_j)$ 表示已知事件 y_j 后所消除的关于事件 x_i 的不确定性，它等于事件 x_i 本身的不确定性 $I(x_i)$ 减去已知事件 y_j 后对 x_i 仍然存在的不确定性 $I(x_i \mid y_j)$。

互信息 $I(x_i;y_j)$ 定义为

$$I(x_i;y_j) = I(x_i) - I(x_i \mid y_j) = \log \frac{p(x_i \mid y_j)}{p(x_i)}$$

3. 平均自信息

平均自信息表示整个信源（用随机变量 X 表示）的平均不确定性，它等于随机变量 X 的每个可能取值的自信息 $I(x_i)$ 的统计平均值。平均自信息量 $H(X)$ 定义为

$$H(X) = E[I(x_i)] = -\sum_{i=1}^{q} p(x_i)\log p(x_i)$$

式中，q 为随机变量 X 的所有可能取值的个数。

4. 离散信源的最大熵

离散信源中各消息等概率出现时熵最大，也称为最大离散熵定理：

$$H(p_1,p_2,\cdots,p_n) \leqslant H\left(\frac{1}{n},\frac{1}{n},\cdots,\frac{1}{n}\right) = \log n$$

式中，n 是随机变量 X 的所有可能取值的个数。

5. 联合熵

联合熵 $H(XY)$ 表示二维随机变量 X、Y 的平均不确定性，它等于联合自信息的统计平均值。联合熵 $H(XY)$ 定义为

$$H(XY) = \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_iy_j)I(x_iy_j) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_iy_j)\log p(x_iy_j)$$

式中，n 和 m 为随机变量 X 和 Y 的所有可能取值的个数。

6. 条件熵

条件熵 $H(Y|X)$ 表示已知随机变量 X 后，对随机变量 Y 仍然存在的平均不确定度。条件熵 $H(Y|X)$ 定义为

$$H(Y|X)=\sum_{i=1}^{n}p(x_i)H(Y|x_i)=-\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_iy_j)\log p(y_j|x_i)$$

式中，$H(Y|x_i)$ 表示已知 $X=x_i$ 的条件下，对随机变量 Y 的平均不确定性。

7. 各类熵之间的关系

$$H(XY)=H(X)+H(Y|X)=H(Y)+H(X|Y)\leqslant H(X)+H(Y)$$

当 X、Y 统计独立时，

$$H(XY)=H(X)+H(Y)$$

8. 平均互信息

平均互信息 $I(X;Y)$ 表示收到一个符号集（用随机变量 Y 表示）后所消除的关于另一个符号集（用随机变量 X 表示）的不确定性，即从 Y 所获得的关于 X 的平均信息量。

平均互信息定义为

$$I(X;Y)=\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_iy_j)I(x_i;y_j)=\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_iy_j)\log\frac{p(x_i|y_j)}{p(x_i)}$$

式中，n 和 m 为随机变量 X 和 Y 的所有可能取值的个数。

9. 平均互信息和各类熵的关系

$$I(X;Y)=H(X)-H(X|Y)=H(Y)-H(Y|X)=H(X)+H(Y)-H(XY)$$

当 X、Y 统计独立时，

$$I(X;Y)=0$$

10. 数据处理定理

如果随机变量 X、Y、Z 构成一个马尔可夫链，则有以下关系成立：

$$I(X;Z)\leqslant I(X;Y)$$

$$I(X;Z)\leqslant I(Y;Z)$$

等号成立的条件是，对于任意的 x，y，z，有 $p(x|yz)=p(x|z)$ 和 $p(z|xy)=p(z|x)$。

数据处理定理中不等式 $I(X;Z)\leqslant I(X;Y)$ 表明，从 Z 所得到的关于 X 的信息量小于等于从 Y 所得到的关于 X 的信息量。如果把 $Y\rightarrow Z$ 视为数据处理系统，那么通过数据处理后，虽然可以满足我们的某种具体要求，但是从信息量来看，处理后会损失一部分信息，最多保持原来获得的信息。也就是说，对接收到的数据 Y 进行处理后，决不会减少关于 X 的不确定性。

11. 熵率

熵率表示离散多符号信源的平均不确定性，它是信源输出的符号序列中，平均每个符号

所携带的信息量。

如果当 $N\to\infty$ 时，极限 $\lim\limits_{N\to\infty}H_N(\boldsymbol{X})$ 存在，则称为熵率，或称为极限熵。熵率定义为

$$H_\infty=\lim_{N\to\infty}H_N(\boldsymbol{X})$$

式中，$H_N(\boldsymbol{X})$ 称为平均符号熵，表示随机变量序列中，对前 N 个随机变量的联合熵的平均：

$$H_N(\boldsymbol{X})=\frac{1}{N}H(X_1X_2\cdots X_N)$$

12. 离散平稳无记忆信源的熵率

多符号信源中最简单的是离散平稳无记忆信源，它的熵率

$$H_\infty=H(X)$$

13. m 阶马尔可夫信源的熵率

离散平稳有记忆信源中比较简单的是记忆长度有限的信源，即马尔可夫信源。如果信源在某时刻发出的符号仅与在此之前发出的 m 个符号有关，则称为 m 阶马尔可夫信源，它的熵率为

$$H_\infty=H(X_{m+1}\mid X_1X_2\cdots X_m)$$

式中，$H(X_{m+1}\mid X_1X_2\cdots X_m)$ 常记为 H_{m+1}，表示已知前面 m 个符号的条件下，输出下一个符号的平均不确定性。

对于离散平稳马尔可夫信源，通常将上述符号的不确定性问题转化为齐次遍历的马尔可夫链的状态转移问题：

$$\begin{aligned}
H_{m+1}&=H(X_{m+1}\mid X_1X_2\cdots X_m)\\
&=E[p(x_{i_{m+1}}\mid x_{i_1}x_{i_2}\cdots x_{i_m})]\\
&=E[p(x_{i_{m+1}}\mid s_i)]\\
&=-\sum_{i=1}^{q^m}\sum_{i_{m+1}=1}^{q}p(s_i)p(x_{i_{m+1}}\mid s_i)\log p(x_{i_{m+1}}\mid s_i)\\
&=\sum_i p(s_i)H(X\mid s_i)\\
&=-\sum_i\sum_j p(s_i)p(s_j\mid s_i)\log p(s_j\mid s_i)
\end{aligned}$$

式中，$p(s_i)$ 是马尔可夫链的状态的平稳分布，$H(X\mid s_i)$ 表示信源处于某一状态 s_i 时发出下一个符号的平均不确定性，$p(s_j\mid s_i)$ 是状态的一步转移概率。

14. 信源剩余度

信源剩余度定义为

$$\gamma=1-\eta=1-\frac{H_\infty}{H_0}=1-\frac{H_\infty}{\log q}$$

式中，H_0 为离散信源的最大熵，H_0-H_∞ 越大，信源的剩余度越大。

15. 连续信源的微分熵

$$h(X)=-\int_R p(x)\log p(x)\,\mathrm{d}x$$

16. 连续信源的最大熵

对于输出信号幅度受限的连续信源，当满足均匀分布时达到最大熵；对于平均功率受限的连续随机变量，当服从高斯分布时具有最大熵。

17. 熵功率

设某连续信源的微分熵为 $h(X)$，则将与它具有相同熵的高斯信源的平均功率 $\overline{P}$ 定义为熵功率，即

$$\overline{P}=\frac{1}{2\pi \mathrm{e}}\mathrm{e}^{2h(X)}$$

假定该连续信源的平均功率为 P，则 $\overline{P}\leqslant P$。熵功率和信源的平均功率相差越大，说明信源的剩余度越大。所以，信源平均功率和熵功率之差 $P-\overline{P}$ 称为连续信源的剩余度。

18. 信道容量

对于给定的信道，即信道转移概率 $p(y_j\mid x_i)$ 固定后，$I(X;Y)$ 是 $p(x_i)$ 的上凸函数。因此对于给定的信道，总存在一种信源（某种输入概率分布），使信道平均传输一个符号接收端获得的信息量最大，也就是说，对于每个固定信道都有一个最大的信息传输率，这个最大的信息传输率即为信道容量，而相应的输入概率分布称为最佳输入分布。信道容量定义为平均互信息对于输入概率分布的最大值：

$$C=\max_{p(x)}\{I(X;Y)\}$$

19. 具有扩展性能的无损信道的信道容量

$$C=\max_{p(x)}\{I(X;Y)\}=\max_{p(x)}H(X)=\log r$$

20. 具有归并性能的无噪信道的信道容量

$$C=\max_{p(x)}\{I(X;Y)\}=\max_{p(x)}H(Y)=\log s$$

21. 具有一一对应关系的无噪无损信道的信道容量

$$C=\max_{p(x)}\{I(X;Y)\}=\max_{p(x)}H(Y)=\max_{p(x)}H(X)=\log s=\log r$$

22. 对称信道的信道容量

$$C=\log s-H(p'_1,p'_2,\cdots,p'_s)$$

式中，$p'_1,p'_2,\cdots,p'_s$ 为信道矩阵中的任一行元素。当输入为等概分布时，达到信道容量。

23. 准对称信道的信道容量

$$C=\log r-\sum_{k=1}^{n}N_k\log M_k-H(p'_1,p'_2,\cdots,p'_s)$$

准对称信道分成 n 个对称的子信道，N_k 是第 k 个子矩阵中行元素之和，M_k 是第 k 个子矩阵中列元素之和。当输入为等概分布时，达到信道容量。

24. 信道容量定理

设有一般离散信道，有 r 个输入符号，s 个输出符号。当且仅当存在常数 C 使式

（1）$I(x_i;Y)=C$，$p(x_i)\neq 0$

（2）$I(x_i;Y)\leqslant C$，$p(x_i)=0$

成立时，$I(X;Y)$ 达到信道容量。其中

$$I(x_i;Y)=\sum_j p(y_j\mid x_i)\log\frac{p(y_j\mid x_i)}{p(y_j)}$$

表示信道输入 x_i 时，所给出的关于输出 Y 的信息量。常数 C 即为所求的信道容量。

25. 离散平稳无记忆信道的 N 次扩展信道的信道容量

$$C^N=NC$$

26. 独立并联信道的信道容量

当 N 个独立并联的信道信道容量都相同时，

$$C_{并}=NC$$

27. 级联信道的信道容量

级联信道的总的信道矩阵等于所级联信道的信道矩阵的乘积。求得级联信道的总的信道矩阵后，级联信道的信道容量就可以用求离散单符号信道的信道容量的方法计算。

28. 波形信道的信道容量

$$C_t=B\log\left(1+\frac{\sigma_X^2}{N_0B}\right) \tag{A.1}$$

这就是著名的香农公式，适用于加性高斯白噪声信道。式中，B 为信道的带宽，N_0 为高斯白噪声的单边功率谱密度。σ_X^2 是输入随机变量 X 的方差，当随机变量 X 的均值为 0 时，即为其平均功率。只有当输入信号为功率受限的高斯白噪声信号时，才能达到该信道容量。

29. 码的分类

- 码
  - 非分组码（树码）
  - 分组码（块码）
    - 奇异码
    - 非奇异码
      - 非唯一可译码
      - 唯一可译码
        - 即时码
        - 非即时码

30. 无失真定长信源编码定理

离散无记忆信源的熵为 $H(S)$，若对信源长为 N 的序列进行定长编码，码符号集中有 r

个码符号，码长为 l，则对于任意小的正数，只要满足 $\frac{l}{N} \geqslant \frac{H(S)+\varepsilon}{\log r}$，则当 N 足够大时，可实现几乎无失真编码，即译码错误概率为任意小。

反之，如果 $\frac{l}{N} \leqslant \frac{H(S)-2\varepsilon}{\log r}$，则不可能实现几乎无失真编码，当 N 足够大时，译码错误概率为 1。

31. Kraft 不等式（McMillan 不等式）

设信源符号集为 $S=\{s_1,s_2,\cdots,s_q\}$，码符号集为 $X=\{x_1,x_2,\cdots,x_r\}$，对信源进行编码，得到的码为 $C=\{w_1,w_2,\cdots,w_q\}$，码长分别为 $l_1,l_2,\cdots,l_q$，即时码（唯一可译码）存在的充要条件是

$$\sum_{i=1}^{q} r^{-l_i} \leqslant 1$$

这称为 Kraft 不等式（McMillan 不等式）。

32. 无失真变长信源编码定理（香农第一定理）

设离散无记忆信源 S 的信源熵为 $H(S)$，它的 N 次扩展信源 $S^N=\{s_1,s_2,\cdots,s_{q^N}\}$，其熵为 $H(S^N)$，用码符号 $X=\{x_1,x_2,\cdots,x_r\}$ 对信源 S^N 进行编码，总可以找到一种唯一可译码，使单个信源符号所需的平均码长满足

$$\frac{H(S)}{\log r} \leqslant \frac{\bar{L}_N}{N} < \frac{H(S)}{\log r} + \frac{1}{N}$$

当 $N \to \infty$ 时，有

$$\frac{\bar{L}_N}{N} = H_r(S)$$

式中，$\bar{L}_N = \sum_{i=1}^{q^N} p(s_i)\lambda_i$，$\lambda_i$ 是扩展信源的信源符号 s_i 所对应的码字长度。

33. 译码规则

设信道的输入符号集 $X=\{x_i\}(i=1,2,\cdots,r)$，输出符号集 $Y=\{y_j\}(j=1,2,\cdots,s)$，若对每个输出符号 y_j，都有一个确定的函数 $F(y_j)$，使 y_j 对应于唯一的一个输入符号 x_i，则称这样的函数为译码规则，记为

$$F(y_j)=x_i \qquad i=1,2,\cdots,r;\ j=1,2,\cdots,s$$

对于有 r 个输入、s 个输出的信道而言，输出 y_j 可以对应 r 个输入中的任何一个，所以译码规则共有 r^s 种。

34. 错误概率

在确定译码规则 $F(y_j)=x_i$ 后，若信道输出端接收到符号 y_j，则一定译成 x_i，如果发送端发送的确实是 x_i，就是正确译码；反之，若发送端发送的不是 x_i，就认为是错误译码。于

是收到符号 y_j 的条件下，译码正确概率为

$$p[F(y_j)\mid y_j]=p(x_i\mid y_j)$$

而译码错误概率为

$$\begin{aligned}p(e\mid y_j)&=1-p[F(y_j)\mid y_j]\\&=1-p(x_i\mid y_j)\end{aligned}$$

式中，e 表示除了 $F(y_j)=x_i$ 以外的所有符号的集合。

35. 平均错误概率

译码后的平均错误概率 P_E 是译码错误概率 $p(e\mid y_j)$ 对 Y 的统计平均值，即

$$\begin{aligned}P_E&=E[p(e\mid y_j)]\\&=\sum_{j=1}^{s}p(y_j)p(e\mid y_j)\end{aligned}$$

它表示平均每接收到一个符号后的译码错误大小。

36. 最大后验概率译码规则

选择译码函数 $F(y_j)=x^*$，使之满足条件

$$p(x^*\mid y_j)\geqslant p(x_i\mid y_j)(x^*\in X)$$

则称为最大后验概率译码规则，又称为最小错误概率准则。对于每个输出符号 $y_j(j=1,2,\cdots,s)$ 均译成与之具有最大后验概率的那个输入符号 x^*，则信道译码的平均错误概率最小。

37. 极大似然译码规则

选择译码函数 $F(y_j)=x^*$，使

$$p(y_j\mid x^*)\geqslant p(y_j\mid x_i)(x^*\in X)$$

称为极大似然译码规则。

当输入符号等概时，最大后验概率译码规则和极大似然译码规则是等价的。

38. 费诺不等式

信道疑义度 $H(X\mid Y)$ 与平均错误概率 P_E 满足以下关系：

$$H(X\mid Y)\leqslant H(P_E)+P_E\log(r-1)$$

费诺不等式表明接收到 Y 后关于 X 的平均不确定性可以分为两部分，第一部分 $H(P_E)$ 是指接收到 Y 后是否产生错误的不确定性，第二部分 $P_E\log(r-1)$ 是已知错误 P_E 发生后，判断是那个输入符号造成错误的最大不确定性，是 $r-1$ 个符号不确定性的最大值与 P_E 的乘积。

39. 汉明距离

长度相同的两个符号序列（码字）$\boldsymbol{x}_i$ 与 $\boldsymbol{y}_j$ 之间的距离是指序列 $\boldsymbol{x}_i$ 与 $\boldsymbol{y}_j$ 对应位置上码元符号不同的位置的个数，称为汉明距离：

$$D(\boldsymbol{x}_i,\boldsymbol{y}_j)=\sum_{k=1}^{n}x_{i_k}\oplus y_{j_k}$$

式中，n 为符号序列的长度。

40. 码的最小距离

码 C 中，任意两个码字的汉明距离的最小值称为该码 C 的最小距离，即

$$D_{\min}=\min\{D(w_i,w_j)\}\qquad w_i\neq w_j,w_i,w_j\in C$$

编码选择码字时，要使码的最小距离越大越好。译码时，则要将接收序列译成与其距离最小的码字，这样得到的平均错误概率最小。

41. 有噪信道编码定理（香农第二定理）

设有一个离散无记忆平稳信道，其信道容量为 C。当待传输的信息率 $R<C$ 时，只要码长 n 足够长，则总存在一种编码，可以使译码错误概率任意小。若 $R>C$，则无论 n 取多大，也找不到一种编码，使译码错误概率 P_E 任意小。

42. 失真函数

设离散无记忆信源 X，经过信道传输后信道输出 Y，对于每对(x_i,y_j)，指定一个非负的函数 $d(x_i,y_j)\geqslant 0(i=1,2,\cdots,r;\ j=1,2,\cdots,s)$，称 $d(x_i,y_j)$ 为单个符号的失真函数或失真度。失真函数表示信源发出一个符号 x_i，而在接收端再现为 y_j 所引起的误差或失真的大小。失真函数通常排列成矩阵形式。

43. 平均失真度

信源的平均失真度表示某个信源通过某个信道传输后失真的大小。

信源的平均失真度定义为

$$\begin{aligned}\overline{D}&=E[d(x_i,y_j)]\\&=\sum_{i=1}^{r}\sum_{j=1}^{s}p(x_iy_j)d(x_i,y_j)\\&=\sum_{i=1}^{r}\sum_{j=1}^{s}p(x_i)p(y_j\mid x_i)d(x_i,y_j)\end{aligned}$$

44. 保真度准则

如果要求信源的平均失真度小于我们所允许的失真 D，即 $\overline{D}\leqslant D$，称为保真度准则。

45. D 失真许可的试验信道

凡满足保真度准则 $\overline{D}\leqslant D$ 的信道称为 D 失真许可的试验信道。D 失真许可的试验信道的集合 B_D 定义为

$$B_D=\{p(y_j\mid x_i):\overline{D}\leqslant D\}\qquad i=1,2,\cdots,r;\ j=1,2,\cdots,s$$

46. 信息率失真函数

对于给定的信源，即信源概率分布 $p(x_i)$ 固定后，$I(X;Y)$ 是信道转移概率 $p(y_j\mid x_i)$ 的下凸函数。因此对于给定的信源，总存在一种信道使 $I(X;Y)$ 达到最小。在满足保真度准则的

D 失真许可的试验信道集合 B_D 中，必然有一个信道 $p(y_j \mid x_i)$ 使 $I(X;Y)$ 达到最小，这个最小值就是信息率失真函数，也称率失真函数。率失真函数 $R(D)$ 定义为

$$R(D) = \min_{p(y_j \mid x_i) \in B_D} I(X;Y)$$

$R(D)$ 是关于 D 的下凸函数，并且在定义域内是严格递减函数。其定义域的下限和上限分别为

$$D_{\min} = \sum_i p(x_i) \min_j d(x_i, y_j)$$

$$D_{\max} = \min_j \sum_i p(x_i) d(x_i, y_j)$$

47. 限失真信源编码定理（香农第三定理）

设 $R(D)$ 是离散无记忆信源的信息率失真函数，并且失真函数为有限值。对于任意的允许失真度 $D \geqslant 0$ 和任意小的正数 $\varepsilon > 0$，当信源序列长度 N 足够长时，一定存在一种编码，其码字个数 $M = e^{N[R(D)+\varepsilon]}$，而编码后的平均失真度 $\overline{D} \leqslant D + \varepsilon$。反过来，若码字个数 $M < e^{NR(D)}$，则必然有 $\overline{D} > D$。

附录 B　习题参考答案

第 1 章　习题参考答案

第 2 章　习题参考答案

第 3 章　习题参考答案

第 4 章　习题参考答案

第 5 章　习题参考答案

第 6 章　习题参考答案

参 考 文 献

[1] 李梅，李亦农，王玉皞．信息论基础教程（第3版）．北京：北京邮电大学出版社，2015.

[2] 傅祖芸．信息论——基础理论与应用．北京：电子工业出版社，2001.

[3] T. M. Cover & J. A. Thomas. Elements of Information Theory. John Wiley & Sons. Inc. , 1991.

[4] （美）R. W. 汉明．编码和信息理论．朱雪龙译．北京：科学出版社，1984.

[5] 朱雪龙．应用信息论基础．北京：清华大学出版社，2001.

[6] 姜丹．信息论与编码．合肥：中国科技大学出版社，2001.

[7] 陈运．信息论与编码．北京：电子工业出版社，2002.

[8] 曹雪虹．信息论与编码．北京：北京邮电大学出版社，2001.

[9] 吴伟陵．信息处理与编码．北京：人民邮电出版社，2003.

[10] 机器学习—4. 大内密探 HMM（隐马尔可夫）围捕赌场老千．ppn029012 的博客．http://blog. csdn. net/ppn029012/article/details/8923501.

[11] 什么是信息增益（Information Gain）．郑来轶的博客．http://blog. sina. com. cn/s/blog_5fc375650100jgxg. html.

[12] 世界第一条量子通信保密干线“京沪干线”将于2016年建成．http://www. mbcom. cn/xwzx/info/12/13/5448.

[13] 最大熵谱估计．http://baike. baidu. com/link?url = y8v4jKjJvc19hXfQBttf0ivK5qY4JDE6SNC F7krJ5pAD6G - vULe6P0pq6XMSGutmz2dG1fwEWk0xs6Jur7XjjywKABxdbG7460hzxX7xUPYuBUpHGdUq4fJXMYAN8EbS8_kBkYbDNxVH3eermYIwkq.

[14] 多传感器信息融合．http://wenku. baidu. com/link?url = 1avkPpat5O4WfltyUsrp1CWnT1XV U07jvWbnPd7ClDABMEWqDwNOrKS - 08oF8ZERGnVyugMuCRHYRrjiPCSSOtXJ_JaLOPdq.

[15] 压缩感知与 Nquist 抽样定理——模拟信息转换（AIC）学习总结．彬彬有礼的专栏．http://blog. csdn. net/jbb0523/article/details/41595535.

[16] （英）David J. C. Mackay. 信息论、推理与学习算法．肖明波，席斌，许芳，王建新译．北京：高等教育出版社，2006.